

Differentially Private Confidence Intervals

Wenxin Du

wendu@reed.edu

Canyon Foot

canafoot@reed.edu

Monica Moniot

mamoniot@reed.edu

Andrew Bray

abray@reed.edu

Adam Groce

agroce@reed.edu

Reed College Mathematics Department
Portland, OR, USA

Abstract

Confidence intervals for the population mean of normally distributed data are some of the most standard statistical outputs one might want from a database. In this work we give practical differentially private algorithms for this task. We provide five algorithms and then compare them to each other and to prior work. We give concrete, experimental analysis of their accuracy and find that our algorithms provide much more accurate confidence intervals than prior work. For example, in one setting (with $\epsilon = 0.1$ and $n = 2782$) our algorithm yields an interval that is only 1/15th the size of the standard set by prior work.

1 Introduction

Estimating the mean of a population is one of the most basic tasks in statistics. A medical researcher who wants to know the average height of an adult male would generally get an estimate by measuring the height of a random sample of people. But when this value is reported, statisticians are usually careful not to just report a single point estimate. They instead include some measure of the *uncertainty* of this estimate. That is, what is the range in which the true average of the population might plausibly fall? This range is given as a *confidence interval*.

In classical statistics, confidence intervals are usually easy to compute. It is often acceptable to assume that continuous variables are (approximately) normally distributed. When estimating the mean of a normally distributed population, given a random sample, the most accurate confidence interval is given by a simple calculation on the sample mean and sample variance. This calculation has been known at least since the codification of confidence intervals in 1937 [14]. However, things become more complicated when confidence intervals are being computed under the constraints of differential privacy.

When computing a confidence interval with differential privacy, there are two sources of randomness to consider.

As in the public setting, there is the sampling variability that arises when selecting a random sample of data. Private algorithms introduce a second source of randomness in order to preserve privacy. There has been surprisingly little work on the computation of private confidence intervals. Karwa and Vadhan [10] study the problem in depth from a theoretical standpoint, finding an algorithm with very good asymptotic performance, but poor practical performance. They state that “designing practical differentially private algorithms for confidence intervals remains an important open problem, whose solution could have wide applicability.” It is precisely this open problem that we are attempting to solve.

Our main contribution is a set of five new, differentially private algorithms that output a confidence interval for the population mean of normally distributed data. We do not assume that the population variance is known. Two of our algorithms use Laplace noise, while three others use an exponential mechanism-based algorithm to report quantiles of the data.

All of our algorithms are experimentally verified to confirm that the resulting intervals are valid. (I.e., they cover the true population mean with the desired frequency.) We also experimentally compare our algorithms to the existing work on this question. We find that our best algorithms consistently outperform prior work, often by large margins. For example, with $\epsilon = 0.1$, a range of $[-32, 32]$, and a sample of size 2782, our best algorithm gives an interval width approximately 2.43 times that obtained without privacy, while the best prior work algorithm gives an interval that is about 37.10 times as wide. This means that the cost of privacy has been reduced by 96%.

We note also that many tasks in private data analysis could probably be made more accurate by assuming some information about the distribution of the underlying data. Our quantile algorithm, for example, when applied to the median, gives a better estimate of the mean of a normal distribution than does the standard Laplace noise-based estimate. Given that many statistical analyses being performed on private data *already*

make these assumptions about its distribution, we see no reason why privacy researchers should not measure utility under these assumptions. (We stress that our privacy guarantees do not depend on any assumptions about the data.)

Our algorithms are all implemented and code is publicly available at <https://github.com/wxindu/dp-conf-int>.

2 Background

Below we first describe differential privacy, a well-established security definition for the private release of statistical queries on sensitive databases. We then discuss the particular sort of query that we study — confidence intervals for the mean of normally distributed data. Finally, we discuss how privacy interacts with the goal of accurate confidence intervals and related prior work.

2.1 Differential Privacy

We imagine an analyst who issues queries to a database of private information. The analyst might be untrusted, or they might be trusted but wish to release the query results publicly. Either way it must be guaranteed that the output of the query protects the privacy of the individuals whose data is contained in the database. Differential privacy [6] formalizes such a guarantee. Intuitively, the guarantee given by differential privacy to an individual is that any output will be roughly equally likely regardless of what data that individual submitted to the database. This implies that no adversary could infer anything about an individual as a result of their participation in the database. (This interpretation is subtle. Interested readers should see [11] or [2] for more discussion.)

To formalize this notion, we first define *neighboring* databases.

Definition 1 (Neighboring Databases). Databases $x, x' \in \mathcal{X}$ are neighbors if one can be transformed to the other by changing the value of a single row x_i .

We can now define differential privacy:

Definition 2 (Differential Privacy). A query f is (ε, δ) -differentially private if for all neighboring $x, x' \in \mathcal{X}$, and for all sets S of possible outputs,

$$\Pr[f(x) \in S] \leq e^\varepsilon \Pr[f(x') \in S] + \delta.$$

This definition, sometimes called *bounded* differential privacy [12], is one of two common variants. The other (*unbounded*) defines neighboring databases as having a row deleted rather than changed. The only significant difference is whether the size of the database must be protected. Our algorithms achieve privacy with $\delta = 0$, so

from here on we state theorems only for the $\delta = 0$ case, though the $\delta > 0$ case is always similar.

The value $\varepsilon > 0$ is considered the *privacy parameter*. Smaller values correspond to less information being revealed about individuals in the database, thus stronger privacy. A value of $\varepsilon = 1$ is fairly high but still meaningfully protects privacy, while $\varepsilon = .01$ is quite low and allows many more queries on the database to be released while still maintaining a strong privacy guarantee, but it also requires queries to be less accurate. The choice of ε is a policy decision.

There are two particularly useful properties of differential privacy that warrant mention. The first, resistance to post-processing, requires that anything computed from private output is itself private. This is a necessary feature of a good privacy definition, but it is also a useful tool that allows the easy construction of private queries.

Theorem 3 (Post-processing [6]). For an ε -differentially private query f , and any function g , the query $g \circ f$ is also ε -differentially private.

Next we give the standard composition theorem, which shows that private queries can be combined in an acceptable way.

Theorem 4 (Composition [6]). If query f_1 is ε_1 -differentially private and query f_2 is ε_2 -differentially private, their composition $f(x) = (f_1(x), f_2(x))$ is $(\varepsilon_1 + \varepsilon_2)$ -differentially private.

Composition allow for the idea of a *privacy budget*, a total ε value that can be divided up as an analyst wishes between any number of queries. It also allows complex queries to be constructed from several smaller, simpler queries.

Finally, we present two general methods for creating private queries. The first is the Laplace mechanism, which adds noise to a query to create a private version. This noise must be proportional the *sensitivity* of the non-private query, defined as the maximum effect a single row can have on the output.

Definition 5 (Global Sensitivity). The sensitivity of a function $f : \mathcal{X} \rightarrow \mathbb{R}$, abbreviated Δf , is defined as

$$\Delta f = \max_{\substack{x, x' \in \mathcal{X} \\ \text{that} \\ \text{are neighbors}}} |f(x) - f(x')|.$$

The noise added is taken from a Laplace distribution, defined below.

Definition 6 (Laplace Distribution). The Laplace Distribution centered at 0 with scale b is the distribution with probability density function:

$$\text{pdf}(z) = \frac{1}{2b} \exp\left(-\frac{|z|}{b}\right).$$

We write $\text{Lap}(b)$ to denote the Laplace distribution with scale b .

This now allows a definition of the Laplace mechanism, the most standard generic technique for privatizing a given query.

Theorem 7 (Laplace mechanism [6]). Given any query $f : \mathcal{X} \rightarrow \mathbb{R}$, the query

$$\tilde{f}(x) = f(x) + L \text{ where } L \sim \text{Lap}\left(\frac{\Delta f}{\varepsilon}\right)$$

is ε -differentially private.

In many cases the Laplace mechanism may not be appropriate. For instance, the output of the query might be categorical rather than numerical, or the query might be numerical but does not have the property that values close to one another have similar usefulness. In these cases, the *exponential mechanism* can be a useful alternative. The exponential mechanism relies on a utility function $u : \mathcal{X} \times R \rightarrow \mathbb{R}$ which assigns a utility to any particular query response, given a particular database. (R is the set of possible query outputs.) The utility function has a sensitivity as well, defined as

$$\Delta u = \max_{\substack{x, x' \in \mathcal{X} \\ r \in R}} |u(x, r) - u(x', r)|.$$

The exponential mechanism selects higher-utility outputs more often, with the probability of a given output increasing exponentially with its utility.

Theorem 8 (Exponential Mechanism [13]). Given a query f and a utility function u , define \tilde{f} such that $\forall x \in \mathcal{X}, \forall r \in R$,

$$\Pr[\tilde{f}(x) = r] \propto \exp\left(\frac{\varepsilon u(x, r)}{2\Delta u}\right).$$

Then \tilde{f} is ε -differentially private.

The exponential mechanism is not necessarily efficiently computable, but it can be for particular queries/utilities.

2.2 Confidence Intervals

In statistical inference, an analyst seeks to use a particular sample of data to infer attributes of the larger population from which it was drawn. One common goal is to estimate a population parameter θ . A confidence interval algorithm takes as input a sample and outputs a range in which the population parameter is likely to fall. Note that we are now considering the database X as a random variable.

Definition 9 (Confidence intervals). Given a database $X = (X_1, \dots, X_n)$ of i.i.d. samples from a population, a confidence interval algorithm c outputs a closed interval $[a, b] \in \mathbb{R}$. An algorithm with confidence level $(1 - \alpha)$, for $\alpha \in [0, 1]$, has the property that

$$\Pr[\theta \in c(X)] \geq 1 - \alpha.$$

The probability in the above definition is traditionally taken over the randomness of the sample X , but it could (and in the private case will) also be taken over the randomness of c , were c to be a randomized algorithm. This probability, $\Pr[\theta \in c(X)]$, is called the *coverage* of the confidence interval algorithm.

Of course, one could construct an interval which is guaranteed to contain the true value by releasing all of \mathbb{R} , but this would not be useful. The goal is to release a small interval, generally measured by the *margin of error (MoE)*, equal to half the interval's width. (I.e., the margin of error for an interval $[a, b]$ is $(b - a)/2$.)

It is acceptable for an algorithm to have coverage *greater* than $1 - \alpha$, but generally when this is the case there is some slack in the algorithm, and the interval can be shrunk to obtain lower average margin of error.

The most well known type of confidence interval, and the kind we focus on in this paper, is a confidence interval for the mean of a normal random variable. In the public setting, this is done with the following algorithm:

Definition 10 (Confidence Intervals for Normally Distributed Data). In the case where the sample comes from a normal distribution, $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, the optimal (smallest MoE) $(1 - \alpha)$ -coverage confidence interval is

$$c(X) = \left[\bar{X} \pm \frac{s}{\sqrt{n}} q_{n-1} \left(\frac{\alpha}{2} \right) \right],$$

where $q_{n-1}(\alpha/2)$ is the $\alpha/2$ quantile of the t -distribution with $n - 1$ degrees of freedom, s is the sample standard deviation, calculated using

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2,$$

and \bar{X} is the sample mean,

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i.$$

2.3 Private Confidence Intervals

Despite the prevalence of confidence intervals in statistical applications, only a handful of papers have attempted to give analysts a way to construct intervals in the private setting. Awan and Slavković [1] derive optimal confidence intervals for binomial proportions, and Sheffet [16] describes confidence intervals for private regression coefficients under certain assumptions.

We are aware of three works that give algorithms for the mean of normally distributed data, those of Karwa and Vadhan [10], D’Orazio et al. [5], and Brawner and Honaker [3]. Because we compare our work to these, we discuss them in more technical detail in Section 3. Additionally, Gaboardi et al. [9] give a method for calculating

confidence intervals in the more restrictive *local* differential privacy setting, in part using the same methods as Karwa and Vadhan.¹

This work on private confidence intervals is directly motivated by attempts to move differential privacy into practice, most specifically the PSI [8] project, which attempts to provide an interface for basic statistics about private data sets used in academic research.

The term “confidence interval” can be used in the private setting in two different ways, which can be confusing. It probably helps to consider the following three kinds of confidence intervals:

Public interval for population mean This is the standard sort of confidence interval thoroughly established in statistics. The goal of the interval is to use the sample data to give an interval estimating the population mean, accounting for the variability induced when selecting the random sample.

Private interval for sample mean This is a common tool in the practice of differential privacy, though rarely discussed in the academic literature. When a mean is reported (e.g., by adding Laplace noise) it is often helpful to give the analyst an understanding of the uncertainty, so a confidence interval can be constructed. This interval is meant to show the uncertainty added by the Laplace noise, so it takes into account only the randomness of the private query. For simple things like the Laplace mechanism, the confidence interval is trivial to construct, though for others it can be very difficult.

Private interval for population mean This is the subject of this paper. Here the goal is to give an interval that will contain the population mean, but to make that algorithm private. That means it must account for the noise of both the random sampling and the private query algorithm.

It is frequently noted in the differential privacy literature that the noise of private mechanisms is known and can therefore be accounted for in statistical analysis, but this accounting has rarely been studied and is actually quite difficult. We believe it is unreasonable to expect people who are not privacy experts to do this accounting and that such an expectation deters the use of differential privacy. Furthermore, some private estimates allow for more accurate noise-accounting than others. For example, one could estimate mean and standard deviation privately and then use the public confidence interval formula, but such a method can fail to accurately guarantee

coverage. Our goal here is to evaluate private algorithms based on the utility of the final, usable output that a practitioner will want to see.

Private confidence interval algorithms vary not just in what margin of error they produce, but also in what assumptions are required. These algorithms, both in prior work and in our work, generally require that the data is known to all come from a given range, i.e., $x_i \in [x_{\min}, x_{\max}]$. (The algorithms work on more general data, but values below x_{\min} are set to x_{\min} before other calculations are performed, and similarly for x_{\max} .) But some algorithms are very sensitive to that range, degrading in accuracy very quickly if the range is overly wide, while others are insensitive to the range, allowing the analyst to give very conservative values.

Finally, we note that both here and in prior work, the assumption that data is normally distributed is only required for the coverage guarantees of the algorithm. Privacy is required to hold in general for any input, regardless of its distribution.

Relationship to hypothesis testing In the public setting, confidence intervals are often discussed interchangeably with hypothesis testing, specifically a t-test. A t-test asks whether a population with a hypothesized mean could plausibly give rise to a sample with the observed sample mean. The probability that such a mean (or a more extreme difference) emerges is a *p*-value. It is typical to check whether $p < 0.05$, and a *p*-value less than 0.05 will occur precisely when the hypothesized mean is outside the resulting confidence interval. As such, any algorithm that produces a confidence interval also produces a hypothesis test and vice versa.

In the private setting, this equivalence no longer holds. Given a hypothesis test, one normally converts to a confidence interval by asking, “At what value does the hypothesis test start rejecting?” But private hypothesis testing algorithms give a *p*-value at one point and cannot be run repeatedly without losing privacy, so there is no way to find where the cutoff for rejecting would be. As a result, the work on private hypothesis testing in this setting [4, 18] does not help us here.

3 Prior Work

Here we describe the three existing works that seek to provide private confidence intervals for the population mean of normally distributed data. We give a very abridged overview of each algorithm; interested readers should refer to the original works for more detail.

3.1 Karwa and Vadhan

The most mathematically sophisticated work comes from Karwa and Vadhan [10]. They give algorithms for both

¹Gaboardi et al. give asymptotic rather than exact analysis of their algorithms, much like Karwa and Vadhan. For the sake of comparison, we implement the work of Karwa and Vadhan and give concrete comparisons. We do not do this for the Gaboardi et al. work, as it is similar in design and in a more restrictive setting, presumably therefore achieving worse performance.

the $\delta = 0$ and the $\delta > 0$ case. In both cases, their algorithm begins by running a private histogram algorithm on the data and uses that to estimate a range for the data. The data is then clamped to that range. Given that truncation, Laplace noise can be added to give a private estimate of the mean and a private but very conservative estimate of variance, which are then used to construct a confidence interval.

This work is impressive and has some very useful results. The margin of error is shown to be asymptotically optimal, and the coverage guarantees hold for finite n , rather than asymptotically². For the $\delta > 0$ case, they require no a priori bounds on the data. In the $\delta = 0$ case, they do require such bounds³, but the accuracy is not highly sensitive to the bounds, so they can be set very conservatively.

However, this work also has serious limitations. While its asymptotic performance is excellent, its practical performance is unacceptable. To quote the paper, “our algorithms are not optimized for practical performance, but rather for asymptotic analysis of the confidence interval length. Initial experiments indicate that alternative approaches (not just tuning of parameters) may be needed to [release] reasonably sized confidence intervals.” These alternative approaches are exactly what we seek to deliver in this paper.

Algorithm 1 Vadhan

Input: $x, \alpha_0, \alpha_1, \alpha_2, \alpha_3, \varepsilon_1, \varepsilon_2, \varepsilon_3$,
 $\bar{s}_{\min}, \bar{s}_{\max}, \tilde{X}_{\min}, \tilde{X}_{\max}$

- 1: $\tilde{X}_{\min}, \tilde{X}_{\max} \leftarrow \text{RANGEFINDER}(x, \alpha_3, \varepsilon_3, \bar{s}_{\min}, \bar{s}_{\max}, \tilde{X}_{\min}, \tilde{X}_{\max})$
- 2: Clamp x by $\tilde{X}_{\min}, \tilde{X}_{\max}$
- 3: $\tilde{X}_{\text{var}} \leftarrow \frac{\tilde{X}_{\max} - \tilde{X}_{\min}}{\varepsilon_1 n}$
- 4: $\tilde{s}_{\text{var}} \leftarrow \frac{(\tilde{X}_{\max} - \tilde{X}_{\min})^2}{\varepsilon_2(n-1)}$
- 5: $\tilde{X} \leftarrow \bar{X} + L_1$, where $L_1 \sim \text{Lap}(\tilde{X}_{\text{var}})$
- 6: $\tilde{s}^2 \leftarrow s^2 + \tilde{s}_{\text{var}} \ln(\frac{1}{2\alpha_2}) + L_2$, where $L_2 \sim \text{Lap}(\tilde{s}_{\text{var}})$
- 7: $MoE \leftarrow \sqrt{\frac{\tilde{s}^2}{n}} q_{t_{n-1}}(1 - \frac{\alpha_0}{2}) + \tilde{X}_{\text{var}} \ln(\frac{1}{\alpha_1})$

Output: $[\tilde{X} - MoE, \tilde{X} + MoE]$

3.2 D’Orazio, Honaker, and King

D’Orazio, Honaker, and King [5] give several private algorithms intended for use in social science research. The only confidence interval they explicitly outline is for the difference of means between two normally distributed variables; however, their method can be easily adapted to produce intervals for a single variable.

²This is somewhat misleading. For low n the algorithm will output \perp instead of a confidence interval. Coverage is correct whenever there is output, but the output is withheld for low n because coverage would not be correct in those cases.

³In fact, they need bounds on both the mean and standard deviation, which is stronger than simply having bounds that hold with high probability on the minimum and maximum data values.

Their algorithm first uses a simple Laplace mechanism query to estimate the sample mean. To get a confidence interval, one needs to compute not just this mean but also an estimate of the sampling variability of the sample mean. To do this, they use an algorithm similar to one of Smith [17]. They first divide the sample into disjoint subsamples and from each calculate such an estimate. This set of estimates S is then fed to a private quantile algorithm to get estimates of the 25th and 75th quantiles. This gives an interquartile range estimate r equal to their difference and a center estimate c equal to their average. The values of S are then truncated to be in $[c - 2r, c + 2r]$. The mean of S is then computed with Laplace noise to ensure privacy.

Once there are estimates of the mean and sampling variability, simulated data can be used to compute an actual confidence interval. We use a similar simulation technique for our algorithms, so we refer the reader to Algorithm 3 and surrounding discussion for more detail.

3.3 Brawner and Honaker

Given a sample mean, statisticians can estimate the variance of that sample mean using bootstrap resampling [7]. Given a database x , a bootstrap sample y of the same size can be computed by randomly sampling from x with replacement. The mean of y is then computed and the process is repeated many times. The distribution of the means of those bootstrapped samples is a good approximation of the sampling distribution of the mean of the original database. For large n , the distribution is known to be asymptotically normal, so the variance of the bootstrapped samples is sufficient to allow the computation of a confidence interval for the population mean.

Despite its prevalence in the practice of statistics, we are familiar with only one (unpublished) paper on private bootstrapping, that of Brawner and Honaker [3]. They give a method that releases k means of bootstrapped samples, each with $1/k^{\text{th}}$ of the privacy budget. These are used to calculate a variance estimate, and they’re also averaged to create an estimate of the sample mean. Crucially, it is shown that the sample mean estimate arrived at this way is just as accurate as one computed directly, but this method avoids the need to allocate part of the budget to the sample mean computation. Given these mean and variance estimates, a confidence interval can be computed. Unfortunately, the variance estimate can often be too low, resulting in unacceptable coverage, but they give a method to conservatively increase the variance estimate and achieve acceptable coverage.

This result is achieved under *zero-concentrated differential privacy* (zCDP) [19]. This privacy is parameterized by ρ , and for any $\delta > 0$ one can convert a guarantee of ρ -zCDP into a guarantee of (ε, δ) -differential privacy.

4 Algorithms

We introduce five algorithms to construct ε -differentially private confidence intervals for the mean of normally distributed data. We start with the simple case, which uses the Laplace Mechanism to produce private estimates of mean and standard deviation. For the second, we modify this method to utilize an alternative dispersion metric that results in a function with lower sensitivity. The remaining three algorithms rely on the exponential mechanism to generative private estimates of quantiles of the data. Each of these methods take a different approach for turning these quantiles into estimates of the center and spread of the sample data, which are then used to construct the confidence interval.

4.1 Noisy Mean and Variance

Our first approach is a direct application of Laplacian noise to the sample mean and variance. The noisy mean and variance are then used to construct the appropriate confidence interval. In this algorithm and several future algorithms, ρ is an allocation parameter that determines the fraction of ε used at various stages of the algorithm. We optimize this value experimentally.

Algorithm 2 Noisy Mean and Variance, NOISYVAR

Input: x, ε, ρ

- 1: $\tilde{x} \leftarrow \bar{x} + L_1$ where $L_1 \sim \text{Lap}\left(\frac{(x_{\max} - x_{\min})}{\rho\varepsilon n}\right)$
- 2: $\tilde{s} \leftarrow \sqrt{\max(0, s^2 + L_2)}$
where $L_2 \sim \text{Lap}\left(\frac{(x_{\max} - x_{\min})^2}{(1-\rho)\varepsilon n}\right)$

Output: \tilde{x}, \tilde{s}

Here x is any database, ε is the privacy parameter, and $0 \leq \rho \leq 1$ is the allocation parameter of ε among queries. We define the sample mean $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$, and the sample variance $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$.

Lemma 11. $\Delta\bar{x} = \frac{(x_{\max} - x_{\min})}{n}$.

Proof. Given $x, x' \in \mathcal{X}$ that are neighbors, the entry that was changed between them can only change value by at most $(x_{\max} - x_{\min})$. As other entries remain unchanged, the sum of all entries can be changed by at most $(x_{\max} - x_{\min})$. When taking the mean of these databases then, the mean can only change value by at most $\frac{(x_{\max} - x_{\min})}{n}$. Thus $\Delta\bar{x} \leq \frac{(x_{\max} - x_{\min})}{n}$. \square

Lemma 12. $\Delta s^2 \leq \frac{(x_{\max} - x_{\min})^2}{n}$.

Proof. This proof modifies Honaker's sensitivity proof for variance estimator $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$. In its place we use the unbiased sample variance $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$. See Appendix A for the full proof. \square

Theorem 13. NOISYVAR is ε -differentially private.

Proof. By composition, it is sufficient to show that \tilde{x} is $\rho\varepsilon$ -differentially private and \tilde{s} is $(1-\rho)\varepsilon$ -differentially private. In step 1 of NOISYVAR, the amount of noise added is exactly $\text{Lap}(\frac{\Delta\bar{x}}{\rho\varepsilon})$, thus \tilde{x} is $\rho\varepsilon$ -differentially private by the properties of the Laplace Mechanism. Similarly, in step 2 of NOISYVAR the amount of noise is $\text{Lap}(\frac{\Delta s^2}{(1-\rho)\varepsilon})$, so $s^2 + L_2$ is $(1-\rho)\varepsilon$ -differentially private and by post-processing, \tilde{s} is $(1-\rho)\varepsilon$ -differentially private. \square

Simulation When generating public confidence intervals, a t -distribution is used to find the critical value and margin of error for a confidence interval. In our case, however, the addition of Laplacian noise to both the standard deviation and the sample mean, renders the t -distribution no longer appropriate. Reed in 2006 [15] introduces the useful Normal-Laplace distribution and provides its cdf and pdf. However, as we fail to invert the cdf to construct a quantile function, the distribution cannot be used for constructing a confidence interval. In its place, we use simulation to construct the reference distribution for the noisy sample mean with standard deviation \tilde{s} . The margin of error is then estimated to be half of the difference between the $\frac{\alpha}{2}$ quantile and the $1 - \frac{\alpha}{2}$ quantile of the simulated reference distribution. Let $q(x, \alpha)$ be a non-private empirical quantile function that outputs the α quantile of sample x .

Algorithm 3 Confidence interval simulation, SIM

Input: $\alpha, \mathcal{A}, x, \varepsilon, nsim$

- 1: $\tilde{x}, \tilde{s} \leftarrow \mathcal{A}(x, \varepsilon)$
- 2: **For** i from 1 to $nsim$ **do**
- 3: $x' \leftarrow x'_0, \dots, x'_n \sim \mathcal{N}(\tilde{x}, \tilde{s}^2)$
- 4: $\tilde{x}'_i \leftarrow \mathcal{A}(x', \varepsilon)$
- 5: $MoE \leftarrow \frac{q(\tilde{x}', 1 - \frac{\alpha}{2}) - q(\tilde{x}', \frac{\alpha}{2})}{2}$

Output: $\tilde{x} - MoE, \tilde{x} + MoE$

The algorithm outputs a $1 - \alpha$ confidence interval through simulation. The input \mathcal{A} can be any algorithm, such as NOISYVAR, that outputs a private estimate of mean and standard deviation when given a database x .

Since SIM only interacts with database x at step 1 through algorithm \mathcal{A} , SIM is a post-processing algorithm and preserves ε -differential privacy. In the following sections we focus only on algorithms which produce private estimates of mean and standard deviation. This allows a general application of SIM to construct confidence intervals.

4.2 Noisy Absolute Deviations

While a private estimate of the standard deviation can be made by adding noise to the naive estimator, previous work has shown that one can increase utility by using an alternative estimator, one with lower sensitivity [20].

Definition 14 (Mean Absolute Deviation). The mean

absolute deviation of the sample x_1, \dots, x_n is

$$\frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|.$$

For normally distributed data x_1, \dots, x_n , the ratio of mean absolute deviation to standard deviation is $\sqrt{\frac{2}{\pi}}$.

The mean absolute deviation has lower sensitivity than s^2 , therefore reducing the amount of noise necessary to maintain privacy. We use $\frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}| \cdot \sqrt{\frac{\pi}{2}}$ to then convert the mean absolute deviation to the sample standard deviation.

Algorithm 4 Noisy absolute deviations, NOISYMAD

Input: x, ε, ρ

- 1: $\tilde{x} \leftarrow \bar{x} + L_1$, where $L_1 \sim \text{Lap}\left(\frac{(x_{\max} - x_{\min})}{\varepsilon_1 n}\right)$
- 2: $\tilde{s} \leftarrow \sqrt{\frac{\pi}{2}} \cdot \max(0, \frac{1}{n} \sum_{i=1}^n |x_i - \tilde{x}| + L_2)$, where $L_2 \sim \text{Lap}\left(\frac{2(x_{\max} - x_{\min})}{(1-\rho)\varepsilon n}\right)$

Output: \tilde{x}, \tilde{s}

The sensitivity of sample standard deviation is $\sqrt{\frac{(x_{\max} - x_{\min})^2}{n}}$ while the sensitivity of the mean absolute deviation is $\frac{2(x_{\max} - x_{\min})}{n}$ which is asymptotically smaller.

Lemma 15. $\Delta \tilde{s} \leq \frac{2(x_{\max} - x_{\min})}{n}$.

Proof. We show in Appendix A that the sensitivity of $\sum_{i=1}^n |x_i - \bar{x}|$ is bounded by $2(x_{\max} - x_{\min})$. Thus the sensitivity of \tilde{s}_0 is bounded by $\frac{2(x_{\max} - x_{\min})}{n}$, and by post-processing, $\Delta \tilde{s} \leq \frac{2(x_{\max} - x_{\min})}{n}$. \square

Theorem 16. NOISYMAD is ε -differentially private.

Proof. It is sufficient to show that \tilde{s} is $(1 - \rho)\varepsilon$ -differentially private, since from the above section we already know that \tilde{x} is $\rho\varepsilon$ -differentially private. In step 2 of NOISYMAD the noise added is $\text{Lap}(\frac{\Delta \tilde{s}}{(1-\rho)\varepsilon})$. Thus \tilde{s} follows the Laplace Mechanism and is $(1 - \rho)\varepsilon$ -differentially private. \square

Using the estimates of mean and standard deviation given by algorithm NOISYMAD, we can construct a private confidence interval using algorithm SIM.

4.3 Exponential Quantiles

The construction of confidence intervals for the mean of normal data requires some measure of the center and spread of the data. The first two methods address this fairly directly, using noisy estimates of the two statistics. A more flexible approach to estimating center and spread is to work with quantiles. A single quantile is a location statistic, telling us the magnitude of a particular part of the distribution (e.g. the 10th quantile tells us where the left tail is; the 50th tells us where the center is). With two quantiles, it becomes possible to estimate the spread of the data.

We introduce three approaches to computing the mean and standard deviation of data using private quantiles. All three methods rely on the algorithm below, which outputs ε -differentially private estimates of any desired sample quantiles using the exponential mechanism. (This algorithm first appears in print in the work of Smith [17], who credits it to McSherry and Talwar [13] and personal correspondence. We know of no published proof of its privacy. Under the unbounded differential privacy definition the proof is almost trivial, but the proof in the bounded case is a bit more complex, so we include it here.)

To concisely explain the algorithm, we rely on the following notation:

- Allow that the values of any given database, x , are sorted: $x_1 \leq \dots \leq x_n$.
- Allow for notational convenience that $x_0 = x_{\min}$ and $x_{n+1} = x_{\max}$.
- Define bins $B_0, \dots, B_n \subseteq [x_{\min}, x_{\max}]$ s.t. $B_i = [x_i, x_{i+1})$.
- Let m indicate the rank of the quantile of interest. (I.e., x_m is the ideal output.)

Since this algorithm uses the exponential mechanism, we must define its utility function. This function is selected such that utility increases as the number of data-points that lie between a potential response and the true quantile decrease. This results in values directly adjacent to the true quantile, x_m , having the highest utility. Let utility function $U_m : \mathcal{X} \times [x_{\min}, x_{\max}] \rightarrow \mathbb{R}$ s.t. $\forall i \in \{0, \dots, n\}$ and for all possible responses $y \in B_i$,

$$U_m(x, y) = U_m(x, B_i) = \begin{cases} i + 1 - m & \text{if } i < m \\ m - i & \text{if } i \geq m \end{cases}.$$

Algorithm 5 Exponential quantile, EXPQ

Input: x, m, ε

- 1: **Define** B_0, \dots, B_n as above
- 2: **For** i from 1 to n **do**
- 3: $p_i \leftarrow |B_i| \cdot \exp(\frac{\varepsilon}{2} U_m(x, B_i))$
- 4: Normalize p_0, \dots, p_n s.t. $\sum_{i=0}^n p_i = 1$
- 5: Sample $i \in [0, n]$ from the distribution defined by p_0, \dots, p_n

Output: $Y \sim \text{Unif}(B_i)$

The range $[x_{\min}, x_{\max}]$ is split into $n + 1$ bins $[x_i, x_{i+1})$ where $i \in \{0, \dots, n\}$. Each bin is assigned a utility score based on its distance from the quantile of interest. Then the exponential mechanism is used to select a bin. The algorithm then randomly outputs a number from the range of the selected bin.

We now show that EXPQ is ε -differentially private.

Lemma 17. Given $x, x' \in \mathcal{X}$ that are neighbors, and their intervals B_0, \dots, B_n , B'_0, \dots, B'_n , if $y \in B_i$ then $y \in B'_{i-1} \cup B'_i \cup B'_{i+1}$.

Proof. Let x^* be a database that has one entry less than both x and x' , and let x_j, x'_k be the removed entries of x, x' such that

$$\begin{aligned} x_0, \dots, x_{j-1}, x_{j+1}, \dots, x_{n+1} &= x_0^*, \dots, x_n^* \\ &= x'_0, \dots, x'_{k-1}, x'_{k+1}, \dots, x'_{n+1}. \end{aligned}$$

So B_0^*, \dots, B_{n-1}^*
 $= B_0, \dots, B_{j-2}, B_{j-1} \cup B_j, B_{j+1}, \dots, B_n$
and B_0^*, \dots, B_{n-1}^*
 $= B'_0, \dots, B'_{k-2}, B'_{k-1} \cup B'_k, B'_{k+1}, \dots, B'_n$.

So $\forall i$, $B_i \subseteq B_{i-1}^* \cup B_i^*$
and $B_i^* \subseteq B'_i \cup B'_{i+1}$
 $\implies B_i \subseteq B'_{i-1} \cup B'_i \cup B'_{i+1}$.

Thus if $y \in B_i$, then $y \in B'_{i-1} \cup B'_i \cup B'_{i+1}$. \square

Lemma 18. $\Delta U_m = 1$.

Proof. Given $x, x' \in \mathcal{X}$ that are neighbors, and $y \in B_i$, allow that x, x' have quantiles of interest x_m, x'_m . By the previous lemma, $y \in B'_{i-1} \cup B'_i \cup B'_{i+1}$. If $i \geq m$, then

$$\begin{aligned} |U_m(x, r) - U_m(x', r)| &= |(m - i) - U_m(x', r)| \\ &\leq |(m - i) - U_m(x', B'_{i\pm1})| \\ &\leq |(m - i) - (m - i \pm 1)| \\ &= |\pm 1| \\ &= 1. \end{aligned}$$

If $i < m$, then

$$\begin{aligned} |U_m(x, r) - U_m(x', r)| &= |(i + 1 - m) - U_m(x', r)| \\ &\leq |(i + 1 - m) - U_m(x', B'_{i\pm1})| \\ &\leq |(i + 1 - m) - (i \pm 1 + 1 - m)| \\ &= |\pm 1| \\ &= 1. \end{aligned}$$

Thus $\Delta U_m = 1$. \square

Theorem 19. EXPQ is ε -differentially private.

Proof. It is sufficient to show that EXPQ follows the exponential mechanism with an output range of $[x_{\min}, x_{\max}]$. Let i, p_0, \dots, p_n be the final values of those same variables in EXPQ.

$$\begin{aligned} \Pr[\text{EXPQ}(x, m, \varepsilon) = y] &= p_i \Pr[Y = y], Y \sim \text{Unif}(B_i) \\ &\propto p_i \frac{1}{|B_i|} \\ &\propto \exp\left(\frac{\varepsilon}{2} U_m(x, B_i)\right) \\ &= \exp\left(\frac{\varepsilon U_m(x, y)}{2\Delta U_m}\right). \end{aligned}$$

Thus EXPQ is ε -differentially private following the exponential mechanism. \square

Theorem 20. EXPQ outputs an unbiased estimator of the median of symmetric data.

Proof. We prove Theorem 20 by showing that the expected output of EXPQ falls into bin B_m or bin B_{m-1} . These bins are adjacent to the median, which implies the expected output will be the median. See the full proof in Appendix A. \square

Since our data are symmetrically distributed, the sample median is an unbiased estimate of the mean. Therefore, the private median produced by EXPQ is an unbiased estimate of the mean. The algorithm is slightly biased when estimating other quantiles, but the bias is practically insignificant for moderately sized n or ε . We discuss this bias in Appendix B.

4.4 Centered Quantiles

The most straightforward application of EXPQ for constructing confidence intervals is to use its median estimate as our mean estimate and use some other quantile as an estimate of standard deviation. Let q_Z be quantile function of the standard normal distribution and b be the choice of quantile. (We will choose b in practice through experimental optimization.)

Algorithm 6 Centered quantiles, CENQ

Input: x, ε, ρ, b

- 1: $\tilde{x} \leftarrow \text{EXPQ}(x, \lfloor \frac{n+1}{2} \rfloor, \rho\varepsilon)$
- 2: $d \leftarrow \text{EXPQ}(x, \lfloor b(n-1) + 1 \rfloor, (1-\rho)\varepsilon)$
- 3: $\tilde{s} \leftarrow \max\left(0, \frac{d-\tilde{x}}{q_Z(b)}\right)$

Output: \tilde{x}, \tilde{s}

Theorem 21. CENQ is ε -differentially private.

Proof. CENQ interacts with the database only through two queries to EXPQ. The first query uses privacy parameter $\rho\varepsilon$ and the second uses $(1-\rho)\varepsilon$. Thus by composition and post-processing, CENQ is ε -differentially private. \square

4.5 Symmetric Quantiles

Here we take a different approach measuring the center of the data. We use algorithm EXPQ to compute two different quantiles an equal distance away from the median. The average of these quantiles is used to estimate the mean while the difference is used to estimate standard deviation.

Theorem 22. SYMQ is ε -differentially private.

The proof proceeds in an analogous manner to that of Thm. 21.

Algorithm 7 Symmetric quantiles, SYMQ

Input: x, ε, b

- 1: $d_1 \leftarrow \text{EXPQ}(x, \lfloor b(n-1) + 1 \rfloor, \frac{\varepsilon}{2})$
- 2: $d_2 \leftarrow \text{EXPQ}(x, \lfloor (1-b)(n-1) + 1 \rfloor, \frac{\varepsilon}{2})$
- 3: $\tilde{x} \leftarrow \frac{d_1+d_2}{2}$
- 4: $\tilde{s} \leftarrow \max\left(0, \frac{d_2-\tilde{x}}{q_Z(1-b)}\right)$

Output: \tilde{x}, \tilde{s}

4.6 Median of deviations

The final approach uses algorithm EXPQ to first compute the median as our estimate of the mean. To estimate standard deviation, we compute the distance between every datapoint and the estimated mean, then we take the median of these distances to estimate standard deviation.

Algorithm 8 Median of deviations, MOD

Input: x, ε, ρ

- 1: $\tilde{x} \leftarrow \text{EXPQ}(x, \lfloor \frac{n+1}{2} \rfloor, \rho\varepsilon)$
- 2: $x' \leftarrow |x_1 - \tilde{x}|, \dots, |x_n - \tilde{x}|$
- 3: $\tilde{s} \leftarrow \frac{\text{EXPQ}(x', \lfloor \frac{n+1}{2} \rfloor, (1-\rho)\varepsilon)}{q_Z(.75)}$

Output: \tilde{x}, \tilde{s}

Theorem 23. MOD is ε -differentially private.

Proof. MOD is another algorithm that composes and post-processes two queries to EXPQ. The interaction in step 2 is private since the database is only being modified element-wise by \tilde{x} , and the only information read from this new database is through the private query to EXPQ at step 3. Thus MOD is ε -differentially private. \square

5 Experimental Results

Because we are focused on concrete performance at low n , rather than asymptotic analysis, we must evaluate our algorithms experimentally. The first thing we must do is experimentally optimize the parameters of each algorithm. Having done that, we must check that they output confidence intervals with the promised coverage. Finally, we must compare them to find the best algorithm(s) and then compare those to prior work.

5.1 Parameter Optimization

The algorithms NOISYVAR, NOISYMAD, CENQ, and MOD all require an ε allocation parameter than determines what proportion of the privacy budget is consumed at different steps of the algorithm. Additionally, CENQ and SYMQ have a parameter b which corresponds to the quantile(s) used to estimate the standard deviation of the database. In all cases, optimization was done experimentally by varying ρ or b . See Appendix E for figures demonstrating the experimental results. In principle, the optimal parameter could be different for different choices

of n , ε , or range, but we found that in all cases we could pick ρ or b values that were roughly optimal in all cases. (In many cases, there was a large region of choices that seemed roughly equally good.) These parameters were fixed at the values given below, and all the following results use these parameter values.

Algorithm	Parameters
NOISYVAR	$\rho = 0.8$
NOISYMAD	$\rho = 0.85$
CENQ	$\rho = 0.5, b = 0.65$
SYMQ	$b = 0.35$
MOD	$\rho = 0.5$

5.2 New Algorithms

Our first experiments sought to determine which of our own algorithms performed best. We compared them with respect to their average *MoE* while varying other parameters.

Figure 1a shows the *MoE* of our algorithms at $\varepsilon = .01$ and $(x_{\min}, x_{\max}) = (-6, 6)$. By a database size of roughly 1000, SYMQ is the clear winner, while for smaller databases NOISYMAD was best. (We see that in general NOISYMAD and NOISYVAR are almost identical in *MoE*, with NOISYMAD consistently having an extremely slight edge.) Figure 7 in Appendix D shows that these findings are generally consistent across choices of ε .⁴ As a rule of thumb, we find that SYMQ is the superior algorithm once $n > 100/\varepsilon$.

Figure 7 also shows results as we vary the $[x_{\min}, x_{\max}]$ range in which the data is bounded. Recall that this is a range given by the analyst, and all data outside the range is moved inside it (set equal to x_{\min} or x_{\max}) before the algorithm is applied. Recall also that some of the prior work goes through great pains to ensure that this range is not needed or can be set very conservatively. The Laplace noise algorithms, NOISYMAD and NOISYVAR, are sensitive to this range, since their noise is proportional to the range. As the range increases their *MoE* increases significantly.

Significantly, we find that our quantile-based algorithms are not sensitive to this range as long as n is not very low. Widening the range only increases the probability that the quantile algorithms pick the most extreme bucket from which to sample. This bucket increases in width proportionately to the range, but its utility decreases exponentially with n . So for reasonably small n , this exponential utility decrease is great enough to make the probability of picking the outermost bucket vanishingly small even when the range is set extremely conservatively. Conveniently, this effect also seems to

⁴There are some exceptions with extremely low ε values and very wide ranges, but our goal here is to find generally useful algorithms, not ones that are marginally less horrible in a weird corner case where everything is bad.

show itself by the time n is approximately $100/\varepsilon$. Therefore we know that in that regime **SYMQ** is an extremely precise algorithm that requires only extremely minimal knowledge of the analyst. Figure 1 shows this for the $\varepsilon = 0.1$ case. Compared to Figure 1, the Laplace noise-based algorithms suffer greatly with a wider range, while the quantile-based algorithms are unaffected. Figure 7 in Appendix D shows the same thing for other parameter settings.

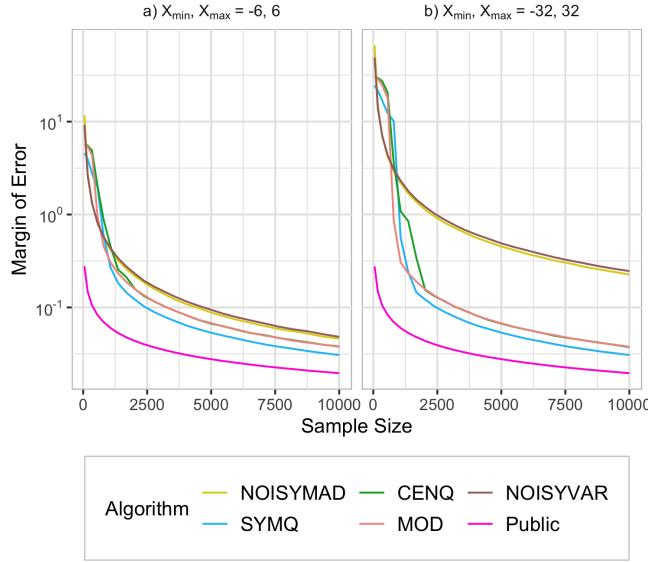


Figure 1: Comparison of our algorithms with respect to their average *MoE* at various database sizes at two different ranges with $\varepsilon = .1$. The distribution of the underlying database was a standard normal, so an x_{\max} of 32 corresponds to 32 standard deviations away from the mean.

We also need to confirm the validity of the algorithms. That is, we must check that their coverage is truly at least $1 - \alpha$. This must be done because the simulated distribution used to calculate the *MoE* is based on an estimated standard deviation for the underlying data. If this estimate is bad enough, it could result in invalid confidence intervals. To do this we run each test many times at many α values and report the percentage of the time that the true mean was included in the resulting interval. Figure 2 shows one example, and Figures 9 and 10 in Appendix D show results at a variety of parameter settings, always with similar (acceptable) results.

What we find is that coverage is generally acceptable. (In these figures, “acceptable” means that the coverage plots never drop below the diagonal.) The one possible exception is the **CENQ** (centered quantiles) algorithm. In some of the experiments it had slightly low coverage for low ε values. Because it is so slight, more work be required to determine for sure whether this was a real issue or just experimental noise. However, **CENQ** is consistently outperformed by **SYMQ** anyway, so it doesn’t

seem worthy of further investigation.

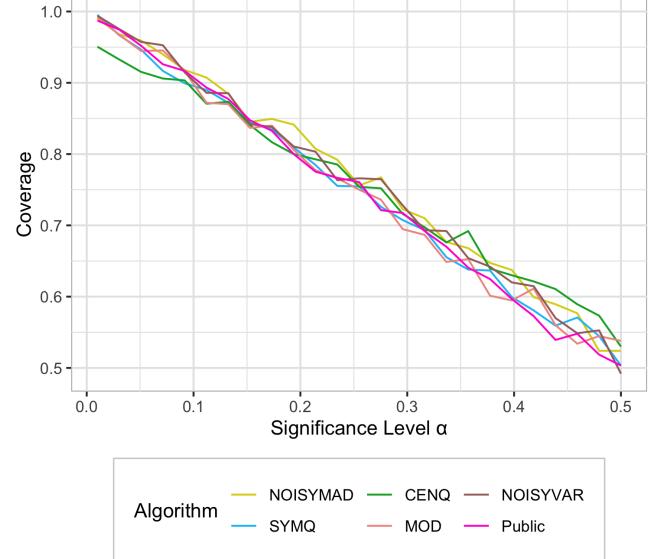


Figure 2: Comparison of our algorithms with respect to coverage when the range is $[-6, 6]$ and $\varepsilon = 0.1$. The underlying distribution of the databases was a standard normal distribution. $n = 1000$

We also check validity in a case where the analyst has not set x_{\min} and x_{\max} so well. In particular, we imagine that the true mean is not centered in the $[x_{\min}, x_{\max}]$ range and that potentially one side of the range is close enough to the true mean to clip a significant number of values. These results can be found in Figure 12 in Appendix D. We find that **NOISYMA** performs fine. We see no problem with **SYMQ** when $n > 100/\varepsilon$. For lower values of n , **SYMQ** does display poor coverage, but at those values it is not the superior algorithm anyway.

5.3 Comparison to Existing Work

We then compare **SYMQ** (our best algorithm for $n > 100/\varepsilon$) and **NOISYMA** (our best algorithm for lower n) to the existing work described in Section 3. We used the same experimental framework as we did before, and part of the results we received are compiled in Figure 3. The full results are shown in Figure 8 in Appendix D.

We varied ε , n , and the data range, and we found that in all cases the lowest *MoE* algorithm was one of ours. In most cases, both **SYMQ** and **NOISYMA** outperformed all prior work. The closest was the work of D’Orazio, Honaker, and King [5].

We also compared coverage between the various algorithms. We again estimated each algorithm’s coverage through simulation, at many different values of α . The experimental results of our two best algorithms and the previously existing algorithms are compiled in Figure 4. The same as before, our algorithms have coverage of roughly $1 - \alpha$, which is ideal. The Karwa and Vadhan [10]

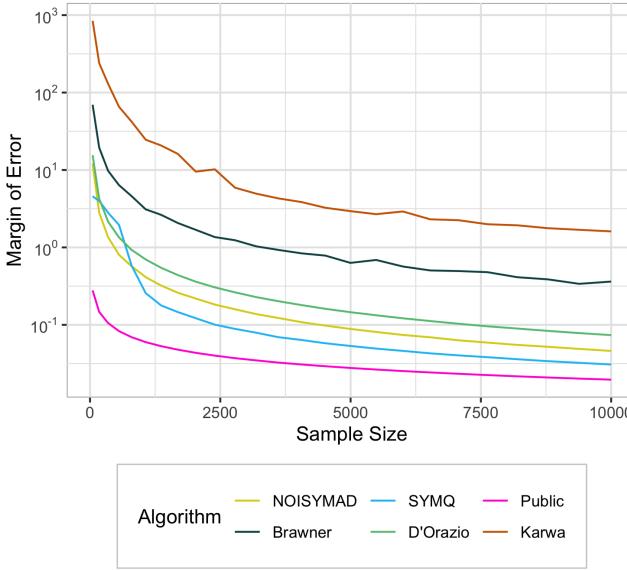


Figure 3: Comparison of our best algorithms to those prior works with respect to their average *MoE* when the range is $[-6, 6]$ and $\varepsilon = 0.1$. The underlying distribution of the databases was a standard normal distribution.

and Brawner and Honaker [3] algorithms have extremely broad coverage, being much more conservative than is necessary. This probably comes from the fact that they use loose upper bounds to set the *MoE*, rather than precise simulation. Figure 11 in Appendix D contains the same experiment run at a variety of range and ε values, all with similar results.

6 Discussion

We have given two practical algorithms for producing confidence intervals for the population mean of normally distributed data. As long as n is somewhat large (at least 100ε), SYMQ performs very well with little drawback. It allows the analyst to set the $[x_{\min}, x_{\max}]$ window extremely conservatively, and the validity is resilient even to a small mistake on the part of the analyst that clips a portion of the data. When n is smaller, NOISYMAD is superior (though in this case the analyst must set $[x_{\min}, x_{\max}]$ a bit more carefully to avoid adding too much noise). It is worth taking a moment to think about why the quantile-based method is so useful here.

Figure 5 shows the distribution of center estimates for normally distributed data. The most accurate (highest-peaked) estimate is of course the sample mean. We also show two private estimates, each at both $\varepsilon = 0.1$ and $\varepsilon = 0.25$. One is the standard Laplace mechanism sample mean estimate. The other is our exponential mechanism-based quantile algorithm, EXPQ, used to find the median. In both cases the quantile algorithm gives a better estimate.

We think this is likely to be part of a larger lesson.

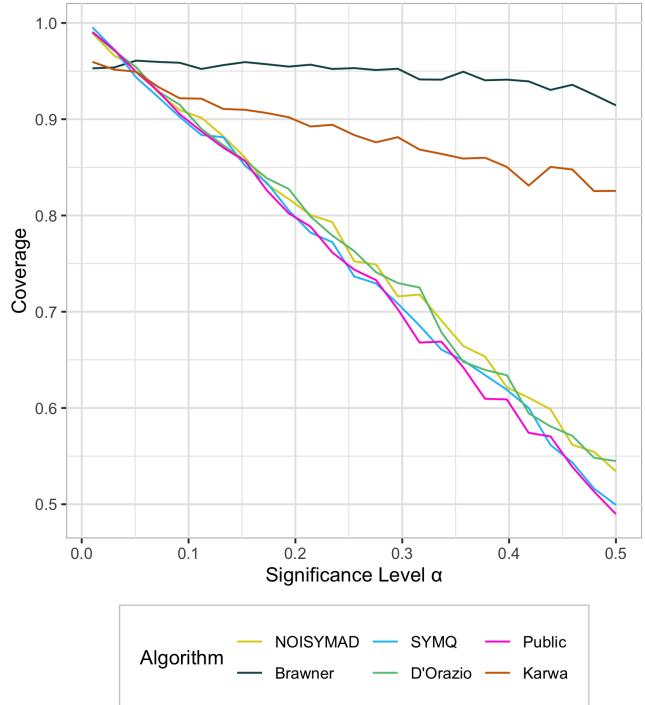


Figure 4: Comparison of our best algorithms to those prior works with respect to coverage when the range is $[-6, 6]$ and $\varepsilon = 0.1$. The underlying distribution of the databases was a standard normal distribution.

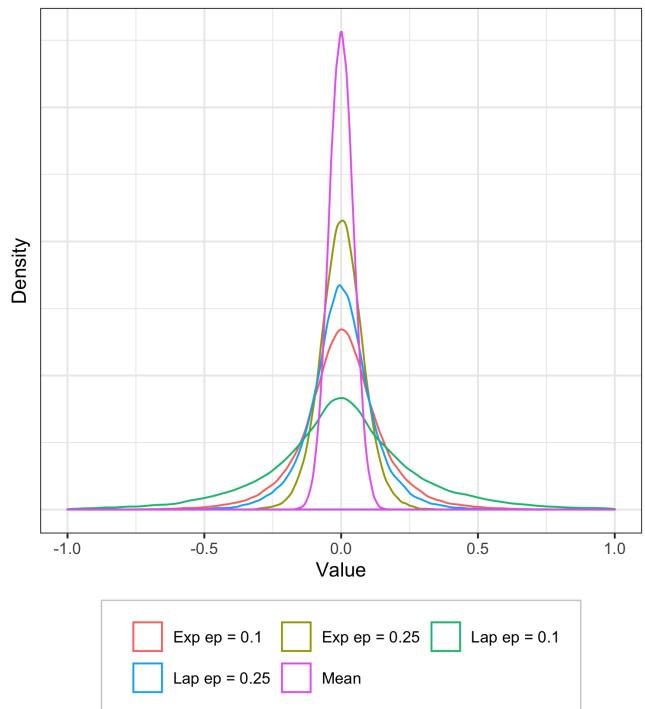


Figure 5: Comparison of the distribution of a sample mean, a private sample mean with Laplace noise, and our exponential mechanism median estimate. $n = 500$, $\varepsilon = 0.1$ or 0.25

The laplacian mechanism is thought of as the “best” algorithm for estimating means, and in some worst-case sense this is true. Medians are thought to be harder to calculate, since their worst-case sensitivity is high, and more complex algorithms are needed. But when data is “nice” or somewhat predictable, the median algorithm is a better estimate of the mean than the “best” mean algorithm. And much of statistics assume some simple properties of the data distribution anyway, so these assumptions are not additional limitations. We think private algorithms should often take into account the likely data distribution. Even when that distribution is not known, using a small portion of the budget for an initial check of possible special cases might often be worthwhile.

In particular, when statistical analysis assumes something about the data, the private version of the analysis should be evaluated under the same assumption.⁵ Given this, it makes sense to design special purpose queries that will be more accurate on particular types of data.

In this work, we’ve given highly practical algorithms for private confidence intervals. It is noteworthy that our best algorithms are quantile-based, relying on an algorithm that is excellent in this setting but that is *not* an ideal way to measure the center of a set of arbitrary data.

7 Conclusion

Our work attempting to find more powerful algorithms for constructing private confidence intervals of the mean of normal data has lead us to several algorithms that perform better at the task than the previously existing work in this area. These algorithms show it is possible in practice to generate small confidence intervals while also providing strong privacy guarantees. Our best algorithm, the symmetric quantiles algorithm **SYMQ**, approaches the public confidence interval quite rapidly for moderately sized n and ϵ . Much of its good performance is due to the exponential quantile algorithm, **EXPQ**, which provides us with more accurate estimates of the mean and standard deviation of a sample than its laplacian noise counterparts. The insensitivity of **EXPQ** to the database range, x_{\min} and x_{\max} also allows our confidence interval algorithms that rely on it to give small intervals despite even the most conservative ranges.

⁵As stated before, this is how *utility* should be measured. Privacy usually should still be a worst-case notion.

References

- [1] Jordan Awan and Aleksandra Slavković. Differentially private uniformly most powerful tests for binomial data. In *Advances in Neural Information Processing Systems*, pages 4208–4218, 2018.
- [2] Raef Bassily, Adam Groce, Jonathan Katz, and Adam Smith. Coupled-worlds privacy: Exploiting adversarial uncertainty in statistical data privacy. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 439–448. IEEE, 2013.
- [3] Thomas Brawner and James Honaker. Bootstrap inference and differential privacy: Standard errors for free. Unpublished Manuscript, 2018.
- [4] Simon Couch, Zeki Kazan, Kaiyan Shi, Andrew Bray, and Adam Groce. Differentially private nonparametric hypothesis testing. *arXiv preprint arXiv:1903.09364*, 2019.
- [5] Vito D’Orazio, James Honaker, and Gary King. Differential privacy for social science inference. *SSRN Electronic Journal*, 01 2015.
- [6] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.
- [7] Bradley Efron. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics*, pages 569–593. Springer, 1992.
- [8] Marco Gaboardi, James Honaker, Gary King, Jack Murtagh, Kobbi Nissim, Jonathan Ullman, and Salil Vadhan. Psi ($\{\Psi\}$): a private data sharing interface. *arXiv preprint arXiv:1609.04340*, 2016.
- [9] Marco Gaboardi, Ryan Rogers, and Or Sheffet. Locally private mean estimation: Z-test and tight confidence intervals. *arXiv preprint arXiv:1810.08054*, 2018.
- [10] Vishesh Karwa and Salil P. Vadhan. Finite sample differentially private confidence intervals. *CoRR*, abs/1711.03908, 2017.
- [11] Shiva Prasad Kasiviswanathan and Adam Smith. A note on differential privacy: Defining resistance to arbitrary side information. *CoRR abs/0803.3946*, 2008.
- [12] Daniel Kifer and Ashwin Machanavajjhala. No free lunch in data privacy. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, pages 193–204. ACM, 2011.
- [13] Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *FOCS*, volume 7, pages 94–103, 2007.
- [14] Jerzy Neyman. X—outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 236(767):333–380, 1937.
- [15] William J. Reed. *The Normal-Laplace Distribution and Its Relatives*, pages 61–74. Birkhäuser Boston, Boston, MA, 2006.
- [16] Or Sheffet. Differentially private ordinary least squares. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3105–3114. JMLR. org, 2017.
- [17] Adam Smith. Privacy-preserving statistical estimation with optimal convergence rates. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pages 813–822. ACM, 2011.
- [18] Eftychia Solea. Differentially private hypothesis testing for normal random variables. 2014.
- [19] Thomas Steinke and Mark Bun. Concentrated differential privacy: Simplifications, extensions, and lower bounds. *Theory of Cryptography Conference*, 10 2016.
- [20] Marika Swanberg, Ira Globus-Harris, Iris Griffith, Anna Ritz, Adam Groce, and Andrew Bray. Improved differentially private analysis of variance. *Proceedings on Privacy Enhancing Technologies*, 2019.

A Unbiased median estimation

Theorem 24. For a sample drawn from a symmetric distribution with symmetric bounds, x_{\min}, x_{\max} , EXPQ is an unbiased estimator of the median.

Let x be the expected value of a database drawn from a symmetric distribution with symmetric bounds.

Let $x_1, \dots, x_n \in [x_{\min}, x_{\max}]$ be database x sorted where $x_1 \leq \dots \leq x_n$.

Allow for notational convenience that $x_0 = x_{\min}$ and $x_{n+1} = x_{\max}$.

Let bins $B_0, \dots, B_n \subseteq [x_{\min}, x_{\max}]$ st $B_i = [x_i, x_{i+1})$.

Let x_m be the median of database x . If n is odd, $m = \frac{n+1}{2}$, if n is even, $m = \frac{n}{2}$.

Let utility function $U_m : \mathcal{X} \times [x_{\min}, x_{\max}] \rightarrow \mathbb{R}$ st $\forall i \in [0, n+1], \forall r \in B_i$,

$$U_m(x, r) = U_m(x, B_i) = \begin{cases} i + 1 - m & \text{if } i < m \\ m - i & \text{if } i \geq m \end{cases}.$$

Lemma 25. Given $\delta \in (0, x_m)$ st $\forall i, x_m \pm \delta \neq x_i$,

$$x_m - \delta \in B_{m-i} \implies x_m + \delta \in B_{m-i+1}.$$

Proof. If n is odd, x_m is the true median.

If n is even, x_m is not the median, but as $n \rightarrow \infty$, it quickly approaches the true median.

Since the distribution of x is symmetric around the median, $x_m - x_{m-i} = x_{m+i} - x_m$.

Given $\delta \in (0, x_m)$ st $\forall i, x_m \pm \delta \neq x_i$, let $i \in [1, m]$ st $x_m - \delta \in B_{m-i}$.

$$\text{So } x_{m-i} < x_m - \delta < x_{m-i+1}$$

$$\implies x_m - x_{m-i+1} < \delta < x_m - x_{m-i}$$

$$\implies x_{m+i-1} - x_m < \delta < x_{m+i} - x_m$$

$$\implies x_{m+i-1} < x_m + \delta < x_{m+i}$$

So $x_m + \delta \in B_{m+i-1}$. □

Proof of Thm. 24. It is sufficient to show that the pdf of the output of $\text{EXPQ}(x, m, \varepsilon)$ is symmetric around the median, x_m .

Since EXPQ follows the exponential mechanism, its pdf must be

$$\phi \cdot \exp\left(\frac{\varepsilon}{2} U_m(x, r)\right)$$

for some normalization constant ϕ .

Given $\delta \in (0, x_m)$, suppose without loss of generality that $\forall i, x_m \pm \delta \neq x_i$.

So $U_m(x, x_m - \delta) = U_m(x, B_{m-i})$ for some $i \in [1, m]$

$$= (m - i) + 1 - m$$

$$= m - (m + i - 1)$$

$$= U_m(x, B_{m+i-1})$$

$$= U_m(x, x_m + \delta).$$

$$\begin{aligned} \text{So } \phi \cdot \exp\left(\frac{\varepsilon}{2} U_m(x, x_m - \delta)\right) \\ = \phi \cdot \exp\left(\frac{\varepsilon}{2} U_m(x, x_m + \delta)\right). \end{aligned}$$

If n is odd, then the pdf of EXPQ is symmetric in the range x_{\min}, x_{\max} around the median.

If n is even, then the pdf of EXPQ is symmetric in the range x_{\min}, x_n around the median. The nonsymmetric part of the pdf, $[x_n, x_{\max}] = B_n$ has probability density $|B_n| \phi \exp\left(\frac{\varepsilon}{2} U_m(x, B_n)\right) = |B_n| \phi \exp\left(\frac{\varepsilon}{4} n\right)$. Since this density approaches 0 rapidly as $n \rightarrow \infty$, the pdf of EXPQ is asymptotically symmetric around the median.

Since the pdf of EXPQ is symmetric around the median, EXPQ is an unbiased estimator of the median. □

B Biased quantile estimation

As we show above, EXPQ is unbiased for estimating the median of normally distributed data. However, this does not hold for other quantiles.

Let X be a random database drawn i.i.d. from a normal distribution.

Let q be the index of the quantile of interest in the database.

So the expected value of $\text{EXPQ}(X, q, \varepsilon)$, an estimate of the $(\frac{q}{n})^{th}$ quantile, is

$$E[\text{EXPQ}(X, q, \varepsilon)] = \int_{x_{\min}}^{x_{\max}} x \cdot f(x) dx$$

Where f is the probability density function of $\text{EXPQ}(X, q, \varepsilon)$. Since f is constant within any bin we can greatly simplify this expression as follows, denoting this constant probability as p_i for the i^{th} bin

$$\begin{aligned} E[\text{EXPQ}(X, q, \varepsilon)] &= \int_{x_{\min}}^{x_{\max}} x \cdot f(x) dx \\ &= \sum_{i=0}^n \int_{X_i}^{X_{i+1}} x \cdot p_i dx \\ &= \sum_{i=0}^n p_i \int_{X_i}^{X_{i+1}} x dx \\ &= \sum_{i=0}^n p_i \frac{X_{i+1}^2 - X_i^2}{2} \\ &= \sum_{i=0}^n p_i (X_{i+1} - X_i) \frac{X_{i+1} + X_i}{2} \end{aligned}$$

Notice that $X_{i+1} - X_i$ is the width of bin i and $\frac{X_{i+1} + X_i}{2}$ is the midpoint of that bin. Although this expression is concise, in practice it is very difficult to work with analytically since the distance between X_i and X_{i+1} will depend on complex order statistics. Given index i , the expected value of X_i is

$$E[X_i] = \int_{-\infty}^{\infty} (i+1) \binom{n}{i} \varphi(x)^{i-1} (1 - \varphi(x))^{n-i} \Phi(x) dx$$

Where $\varphi(x)$ is the pdf of the normal distribution of X , and $\Phi(x)$ is the corresponding cdf. It is easy to see how unwieldy this expression will make deriving analytical results for our estimator. For this reason, we choose to implement this function and plot the bias in our estimator empirically for various values of n and ε .

Theorem 26 (Variance Sensitivity following Honaker).

$$\Delta s^2 = \frac{(x_{\max} - x_{\min})^2}{n}.$$

Proof. Let x, x' be two neighboring datasets which only differ at the j th row. For simpler notation, we let $\bar{x}_{-j} = \frac{1}{n-1} \sum_{i \neq j} x_i$. Then the sample variance can be rewritten as

$$\begin{aligned} s^2 &= \frac{1}{n-1} \sum_{i=1}^n (\bar{x} - x_i)^2 \\ &= \frac{n}{n-1} \bar{x}^2 - 2\bar{x} \frac{1}{n-1} \sum_{i=1}^n x_i + \frac{1}{n-1} \sum_{i=1}^n x_i^2 \\ &= \frac{n}{n-1} \bar{x}^2 - 2\bar{x} \frac{n}{n-1} \bar{x} + \frac{1}{n-1} \sum_{i=1}^n x_i^2 \\ &= \frac{1}{n-1} \sum_{i=1}^n x_i^2 - \frac{n}{n-1} \bar{x}^2 \\ &= \frac{1}{n-1} \sum_{i=1}^n x_i^2 - \frac{1}{n(n-1)} \left(\sum_{i=1}^n x_i \right)^2 \\ &= \frac{1}{n-1} \left(x_j^2 + \sum_{i \neq j} x_i^2 \right) - \\ &\quad \frac{1}{n(n-1)} \left(x_j + \sum_{i \neq j} x_i \right)^2 \\ &= \frac{1}{n-1} \left(x_j^2 + \sum_{i \neq j} x_i^2 \right) - \\ &\quad \frac{1}{n(n-1)} \left[x_j^2 + 2x_j \sum_{i \neq j} x_i + \left(\sum_{i \neq j} x_i \right)^2 \right] \\ &= \frac{1}{n-1} \left[\sum_{i \neq j} x_i^2 - \frac{1}{n} \left(\sum_{i \neq j} x_i \right)^2 \right] + \\ &\quad \frac{1}{n(n-1)} \left[(n-1)x_j^2 - 2x_j \sum_{i \neq j} x_i \right] \\ &= \left[\frac{1}{n-1} \sum_{i \neq j} x_i^2 - \frac{1}{n(n-1)} \left(\sum_{i \neq j} x_i \right)^2 \right] + \\ &\quad \frac{n-1}{n(n-1)} \left(x_j^2 - 2x_j \bar{x}_{-j} \right) \\ &= \left[\frac{1}{n-1} \sum_{i \neq j} x_i^2 - \frac{1}{n(n-1)} \left(\sum_{i \neq j} x_i \right)^2 \right] + \\ &\quad \frac{1}{n} \left(x_j^2 - 2x_j \bar{x}_{-j} \right) \end{aligned}$$

Thus changing x_j would only affect the latter term $\frac{1}{n} (x_j^2 - 2x_j \bar{x}_{-j})$. The difference between variance of x and x' would then be:

$$s^2(x) - s^2(x') = \frac{1}{n} [x_j^2 - x'_j^2 - 2(x_j - x'_j) \bar{x}_{-j}]$$

Now consider the partial derivative of the variance function with respect to the j th observation:

$$\begin{aligned} \frac{\partial s^2}{\partial x_j} &= \frac{2}{n} (x_j - \bar{x}_{-j}) \\ \frac{\partial^2 s^2}{\partial x_j^2} &= \frac{2}{n} > 0 \end{aligned}$$

And thus among all possible values x_j can take, the variance s^2 is minimized when $x_j = \bar{x}_{-j}$ and maximized when x_j is equal to the bound that's farthest from \bar{x}_{-j} . Thus the sensitivity bound of variance can be found to be:

$$\begin{aligned} \Delta s^2 &= \max_{x_j, \bar{x}_{-j}} [s^2(x_j) - s^2(x'_j)] \\ &= \max_{x_j, \bar{x}_{-j}} \frac{1}{n} [x_j^2 - x'_j^2 - 2(x_j - x'_j) \bar{x}_{-j}] \\ &= \max_{x_j, \bar{x}_{-j}} \left[\frac{1}{n} (x_j^2 - \bar{x}_{-j}^2 - 2(x_j - \bar{x}_{-j}) \bar{x}_{-j}) \right] \\ &= \max_{x_j, \bar{x}_{-j}} \left[\frac{1}{n} (x_j^2 - 2x_j \bar{x}_{-j} + \bar{x}_{-j}^2) \right] \\ &= \max_{x_j, \bar{x}_{-j}} \left[\frac{1}{n} (x_j - \bar{x}_{-j})^2 \right] \\ &= \frac{1}{n} (x_{\max} - x_{\min})^2 \end{aligned}$$

□

Theorem 27. Bounding Sensitivity of $f(x) = \sum_{i=1}^n |x_i - \bar{x}|$.

Proof. Consider an arbitrary change in a particular database value x_j , to a new value x'_j . Setting $d = x'_j - x_j$, we notice that the altered mean, $\bar{x}' = \bar{x} + \frac{d}{n}$. Now, we bound the sensitivity as follows,

$$\begin{aligned} \Delta f &= \max_{x, x'} \text{neighbors} |f(x) - f(x')| \\ &= \max_{x, x'} \text{neighbors} \left| \sum_{i \neq j} |x_i - \bar{x}| - \sum_{i \neq j} |x_i - \bar{x}'| \right. \\ &\quad \left. + |x_j - \bar{x}| - |x'_j - \bar{x}'| \right| \\ &\leq \max_{x, x'} \text{neighbors} \left(\left| \sum_{i \neq j} |x_i - \bar{x}| - \sum_{i \neq j} |x_i - \bar{x}'| \right| \right. \\ &\quad \left. + ||x_j - \bar{x}| - |x'_j - \bar{x}'|| \right) \end{aligned}$$

Considering the cases $\sum_{i \neq j} |x_i - \bar{x}| - \sum_{i \neq j} |x_i - \bar{x}'|$ and

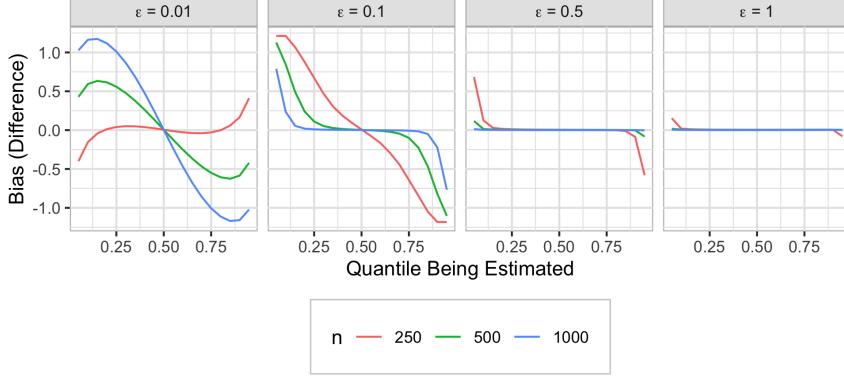


Figure 6: The bias (As difference between true and expected value) of our estimator given different values of n and ϵ

$|x_j - \bar{x}| - |x'_j - \bar{x}'|$ separately, we have

$$\begin{aligned} & \sum_{i \neq j} |x_i - \bar{x}| - \sum_{i \neq j} |x_i - \bar{x}'| \\ &= \sum_{i \neq j} |x_i - \bar{x}| - \sum_{i \neq j} |x_i - (\bar{x} + \frac{d}{n})| \\ &\leq \sum_{i \neq j} |x_i - \bar{x}| - \sum_{i \neq j} |x_i - \bar{x}| + \sum_{i \neq j} |\frac{d}{n}| \\ &= \left| \frac{d(n-1)}{n} \right|, \end{aligned}$$

and

$$\begin{aligned} & |x_j - \bar{x}| - |x'_j - \bar{x}'| \\ &= |x_j - \bar{x}| - |x_j + d - (\bar{x} + \frac{d}{n})| \\ &\leq |x_j - \bar{x}| - |x_j - \bar{x}| + |d| + \left| \frac{d}{n} \right| \\ &= |d| + \left| \frac{d}{n} \right|. \end{aligned}$$

So, putting the two together, we have

$$\begin{aligned} \Delta f &\leq \max_d \left(|d| + \left| \frac{d}{n} \right| + \left| \frac{d(n-1)}{n} \right| \right) \\ &= \left(1 + \frac{1}{n} + \frac{n-1}{n} \right) \max_d |d| \\ &= 2 \max_d |d|. \end{aligned}$$

Since $|d| \leq (x_{\max} - x_{\min})$, $\Delta f \leq 2(x_{\max} - x_{\min})$. \square

C Details on D’Orazio and Honaker’s algorithm

D’Orazio, Honaker, and King [5] describe a method for calculating the standard error for a private estimate of the difference in means between two normally distributed random variables. This is a different case than we are

considering. However, the difference between two normally distributed random variables is itself normally distributed, so their method can be adopted easily enough to our case. Making this change does mean we had to slightly adapt some aspects of the algorithm, and for this reason we have reproduced the exact algorithm we used below.

A few changes and implementation details of note:

1. Although the algorithm used EXPQ to estimate the first and third quartiles of the subsampled estimates of standard error, they did not give any details about how the upper bound given to EXPQ was determined. We felt that since this was a bound on standard error rather than standard deviation, it would need to depend on the size of the database provided. To get around this, we pass a bound for the actual standard deviation, sd_{\max} and divide this by \sqrt{n} to place a bound on the true standard error, se_{\max} . However, this bound is likely not conservative enough because the standard errors will follow their own sampling distribution based on the size of the subsets. To address this, we add two standard deviations of the standard error calculated on the M subsets to se_{\max} . The standard error of the standard deviation calculated on each subset is approximately $\frac{sd_{\max}}{\sqrt{2 \frac{M}{n}}}$. After rearranging terms and scaling by $\frac{1}{\sqrt{n}}$, we get that the standard error on our estimates of standard error is $\frac{sd_{\max} \cdot \sqrt{M}}{n\sqrt{2}}$. We then added three times this value to our bound on the true standard error to get the value we pass to EXPQ.

2. The paper also did not give any discussion of how to select the number of subsets on which to calculate the standard deviation. Smith 2011 [17] gave a heuristic of \sqrt{n} as the number of subsets for a similar algorithm, but we found this to give too few groups. Instead, we empirically optimized the group size at various levels of n and interpolated to approximate

the best subsample size for a given database. We found that for all the sample sizes we tried, the best results occurred when the size of each subsample was 2.

3. In the original paper, the Laplace noise added to the winzorized mean has scale parameter $\frac{|u-l|}{2\varepsilon M}$. We believe this is an error and that the value should be $\frac{2|u-l|}{\varepsilon M}$. This is because half of the ε budget is consumed by the quartile estimates and so the sensitivity $\frac{|u-l|}{M}$ should be divided by the remaining $\frac{\varepsilon}{2}$.

Algorithm 9 Construct D’Orazio Mean and SD, ORA

Input: $x, \varepsilon, M, x_{min}, x_{max}, sd_{max}$

- 1: $\tilde{x} \leftarrow \bar{x} + L_1$, where $L_1 \sim \text{Lap}\left(\frac{2(x_{max}-x_{min})}{\varepsilon n}\right)$
- 2: $se_{max} \leftarrow \frac{sd_{max}}{\sqrt{n}}$
- 3: Divide dataset x into M subsets m_1, m_2, \dots, m_M .
- 4: **For** $i \leftarrow 1, M$ **do**
- 5: $s_i \leftarrow \frac{sd(m_i)}{\sqrt{n}}$
- 6: $S \leftarrow s_1, s_2, \dots, s_M$
- 7: $a \leftarrow \text{EXPQ}\left(S, \frac{1}{4}, \frac{\varepsilon}{4}, 0, se_{max} + 2 \cdot \frac{sd_{max} \cdot \sqrt{M}}{\sqrt{2 \cdot n^2}}\right)$
- 8: $b \leftarrow \text{EXPQ}\left(S, \frac{3}{4}, \frac{\varepsilon}{4}, 0, se_{max} + 2 \cdot \frac{sd_{max} \cdot \sqrt{M}}{\sqrt{2 \cdot n^2}}\right)$
- 9: $\mu \leftarrow \frac{a+b}{2}$
- 10: $IQR \leftarrow |a - b|$
- 11: $u \leftarrow \mu + 2IQR$
- 12: $l \leftarrow \mu - 2IQR$
- 13: **For** $i \leftarrow 1, M$ **do**
- 14: $s_i \leftarrow \begin{cases} u & \text{if } s_i > u \\ s_i & \text{if } l < s_i < u \\ l & \text{if } s_i < l \end{cases}$
- 15: $w \leftarrow \frac{1}{M} \sum_{i=1}^M s_i$
- 16: $\tilde{s} \leftarrow w + L_2$, $L_2 \sim \text{Lap}\left(\frac{2|u-l|}{\varepsilon M}\right)$

Output: \tilde{x}, \tilde{s}

D Detailed experimental results

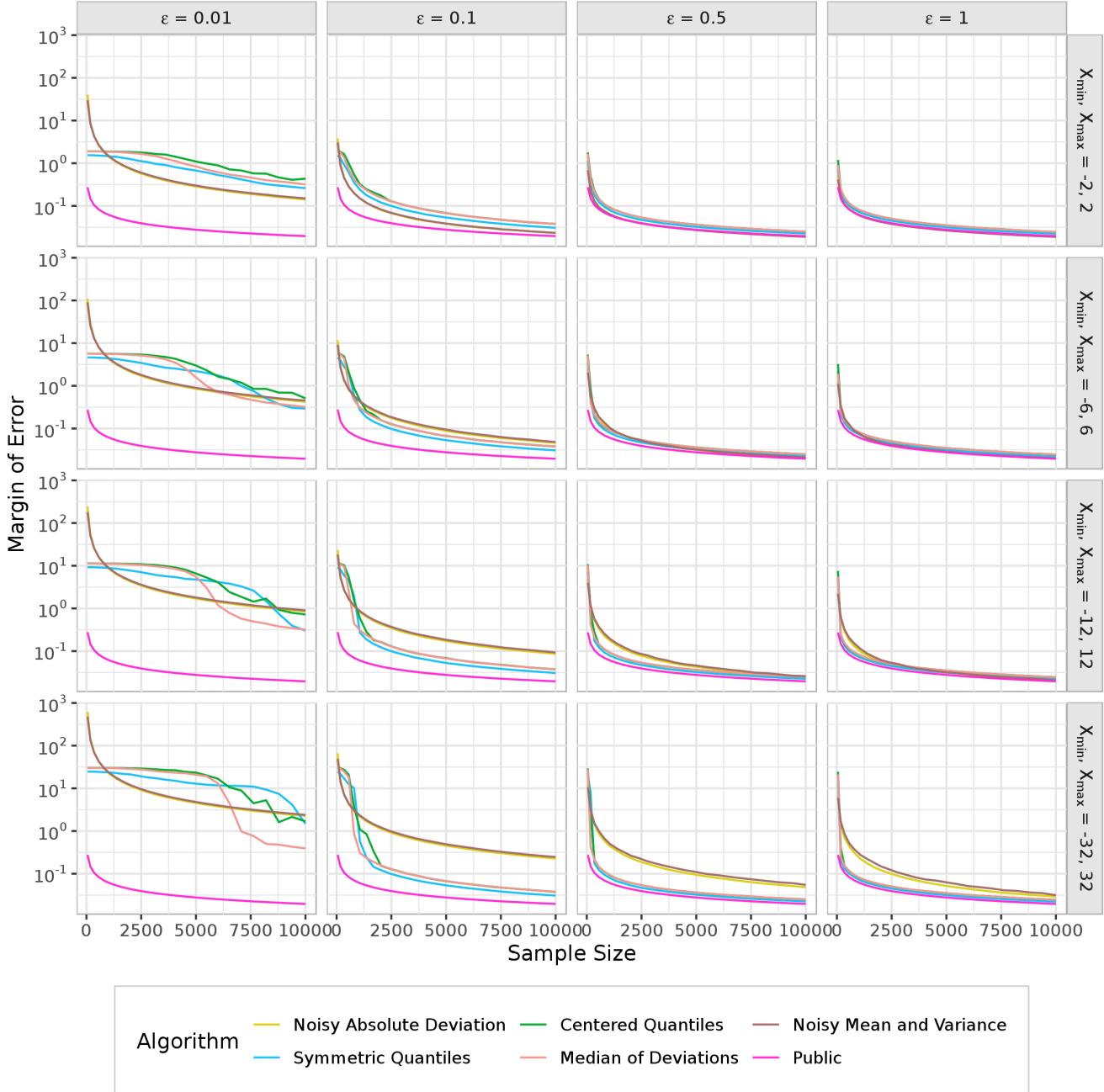


Figure 7: Comparison of our algorithms with respect to their average *MoE* at various database sizes, ε values , and ranges.

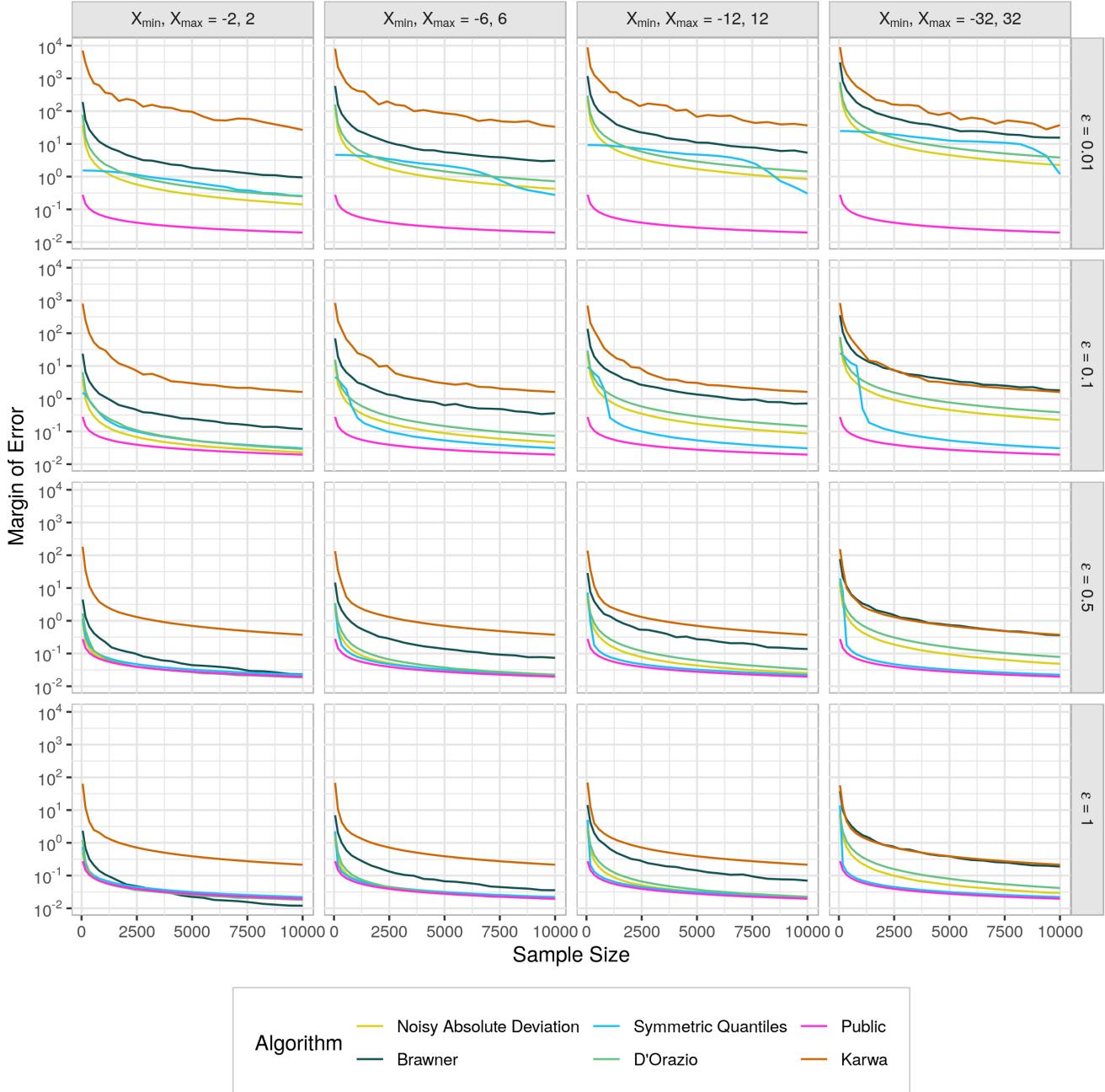


Figure 8: Our best algorithms compared to prior work with respect to their average MoE at various databases sizes, ϵ values and ranges.

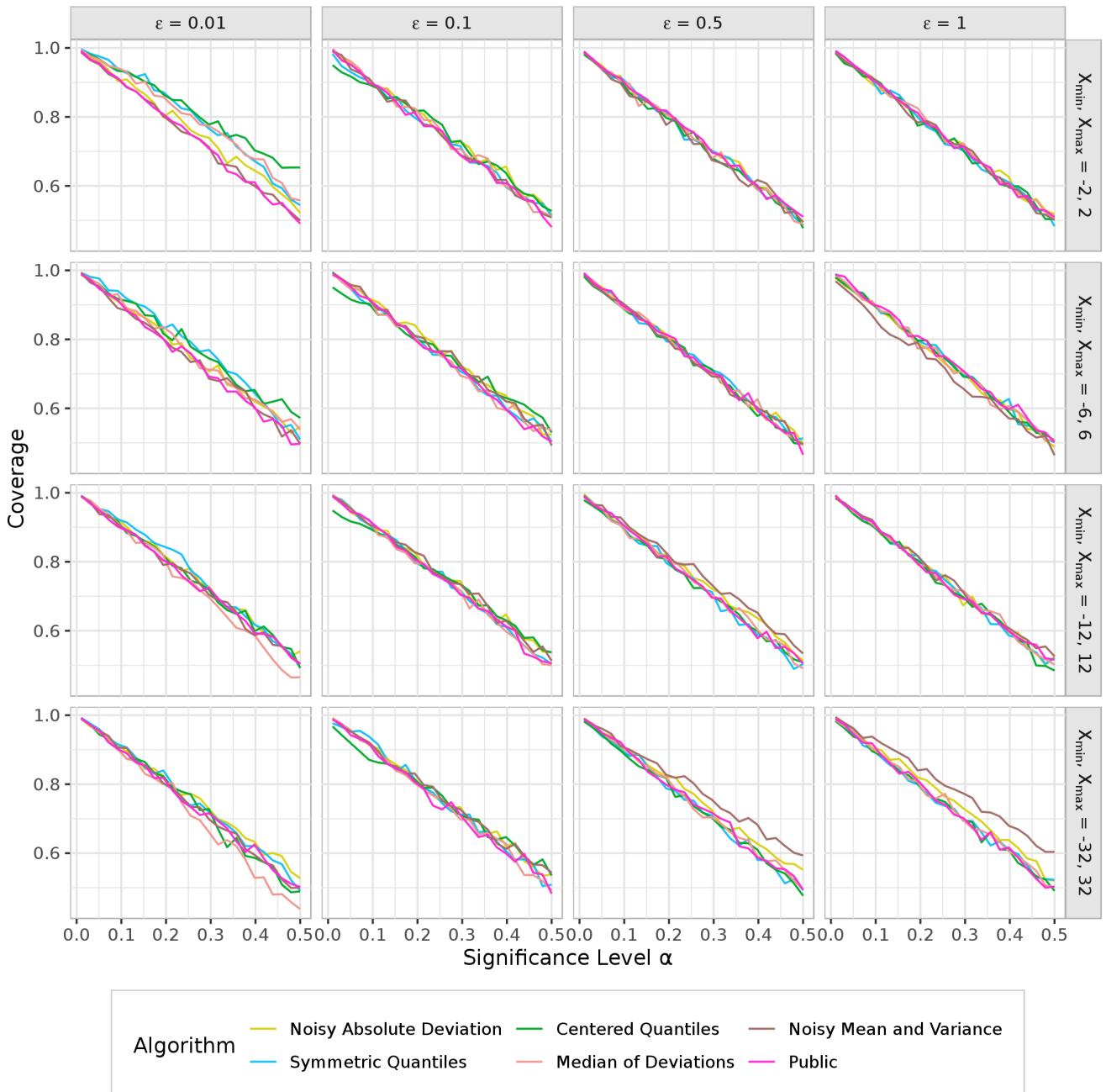


Figure 9: The coverage of our algorithms at various significance levels, ε values and ranges. $n = 1000$

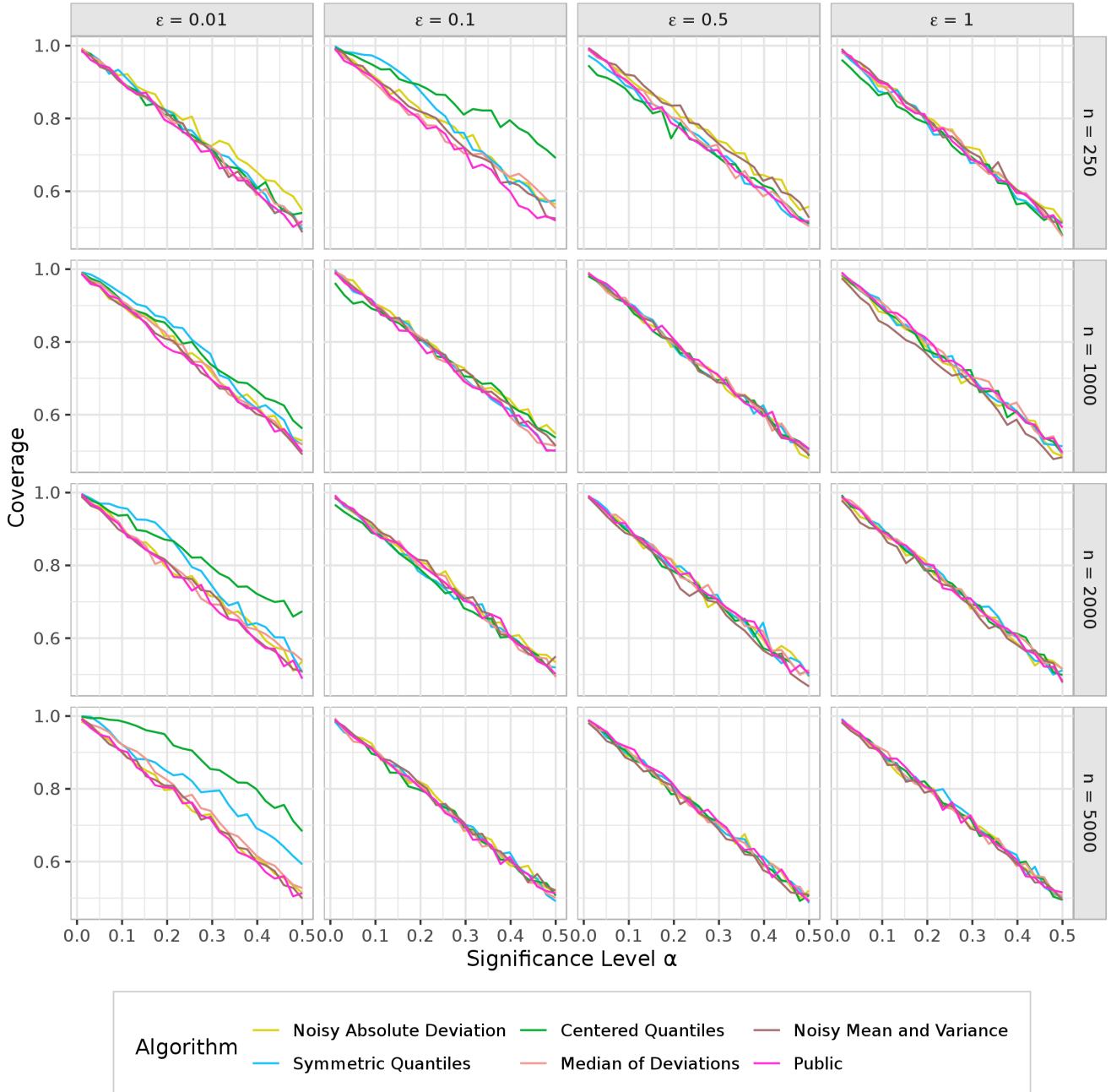


Figure 10: The coverage of our algorithms at various significance levels, ε values and sample sizes.

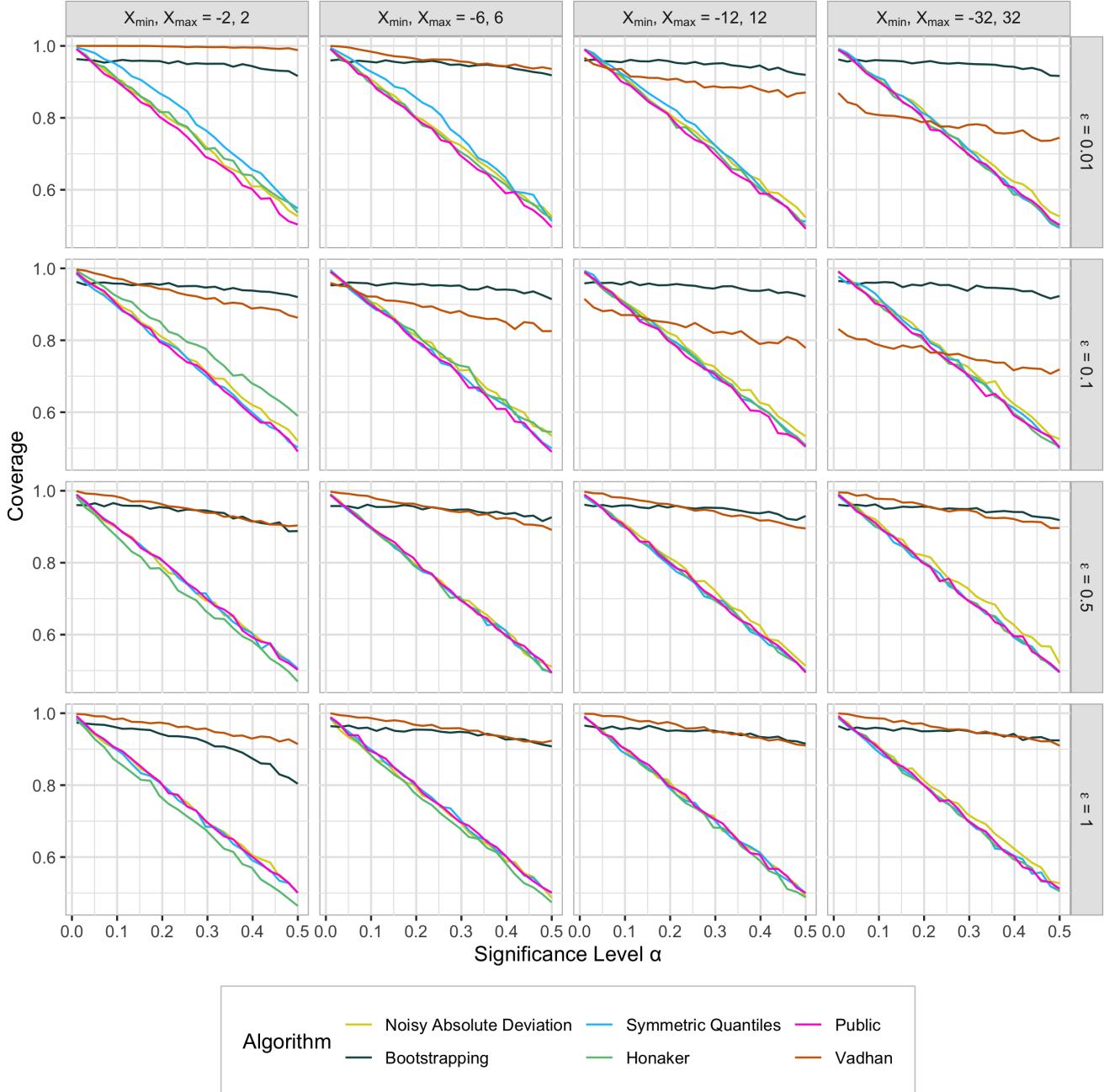


Figure 11: Our best algorithms compared to prior work with respect to their coverage at various significance level, ϵ values and ranges.

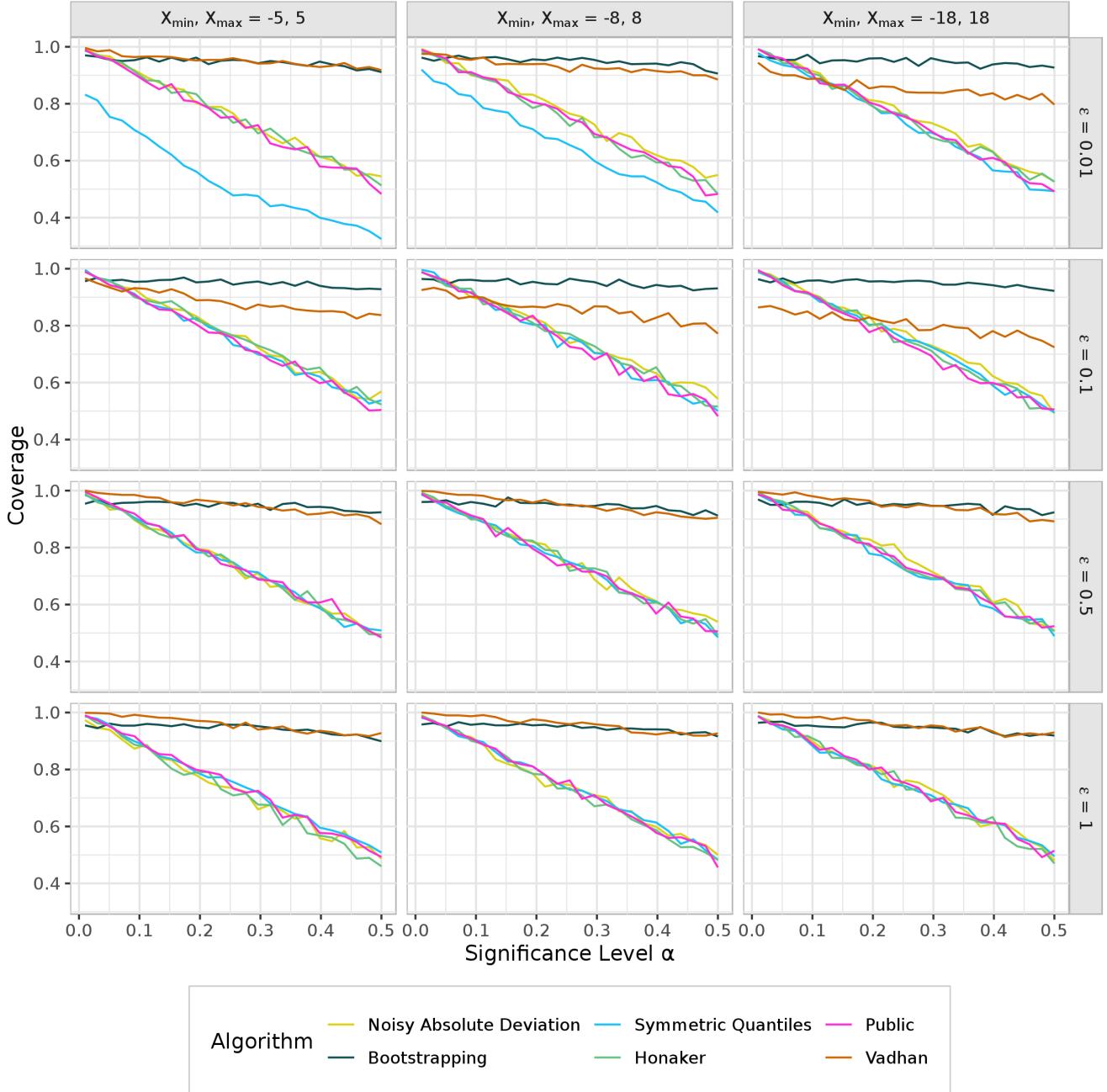


Figure 12: The coverage of our algorithms at various significance levels, ϵ values and sample sizes. Here the true mean is 3 to test an off-center case.

E Parameter Allocation and Optimization

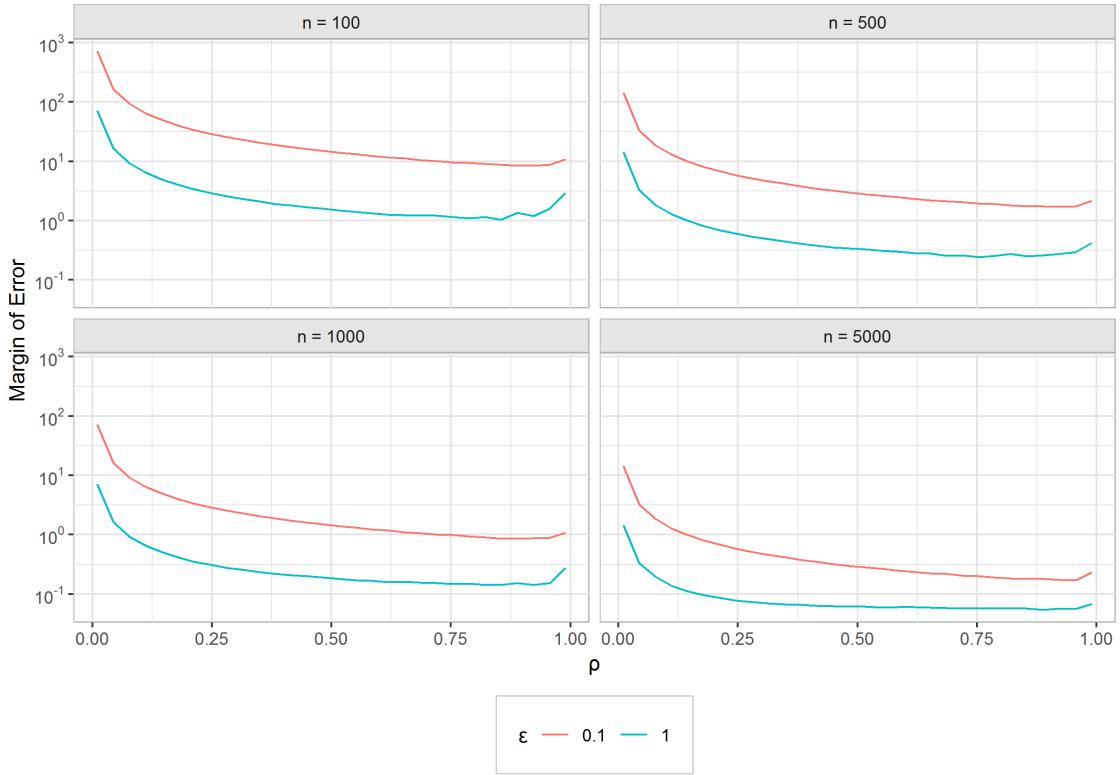


Figure 13: The performance of NOISYVAR at various database sizes varying ε -allocation. NOISYVAR tends to perform the best when ρ is in the range of approximately $(0.75, 0.85)$, where the MoE of confidence intervals are minimized for all database sizes. We thus choose $\rho = 0.8$ as the optimized ε -allocation for NOISYVAR.

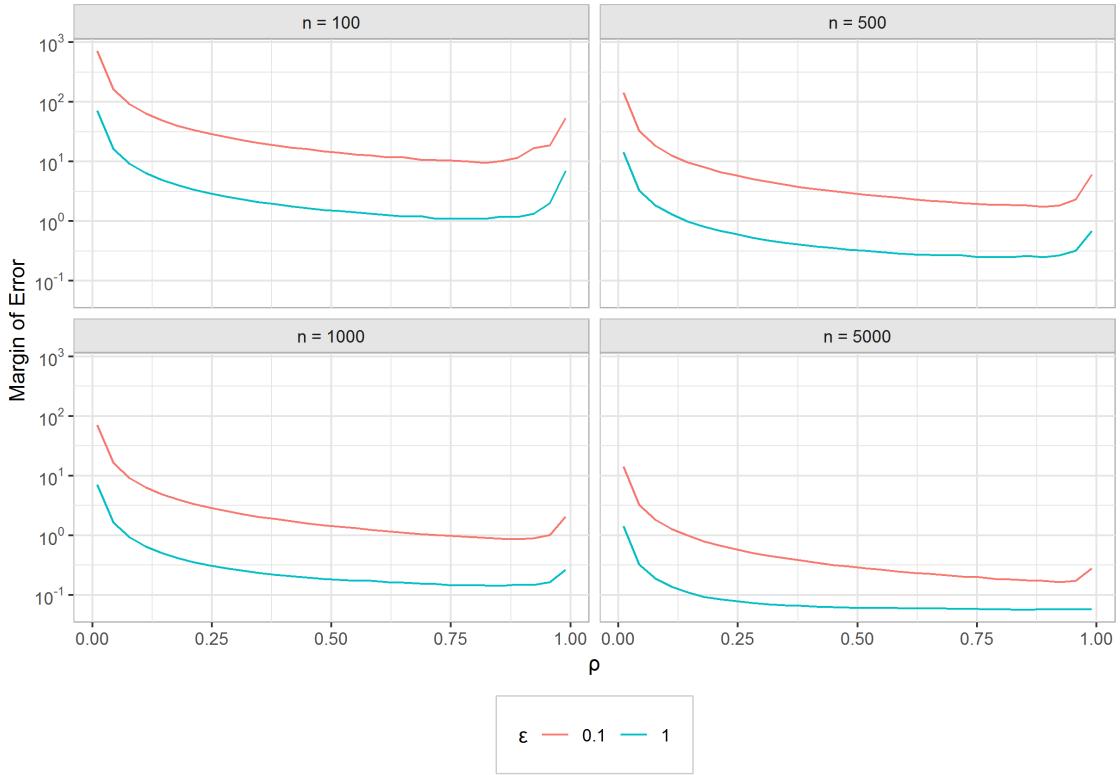


Figure 14: The performance of NOISYMAD at various database sizes varying ε -allocation. NOISYMAD tends to perform the best when ρ is in the range of approximately $(0.75, 0.88)$, where the MoE of confidence intervals are minimized for all database sizes. We thus choose $\rho = 0.85$ as the optimized ε -allocation for NOISYMAD.

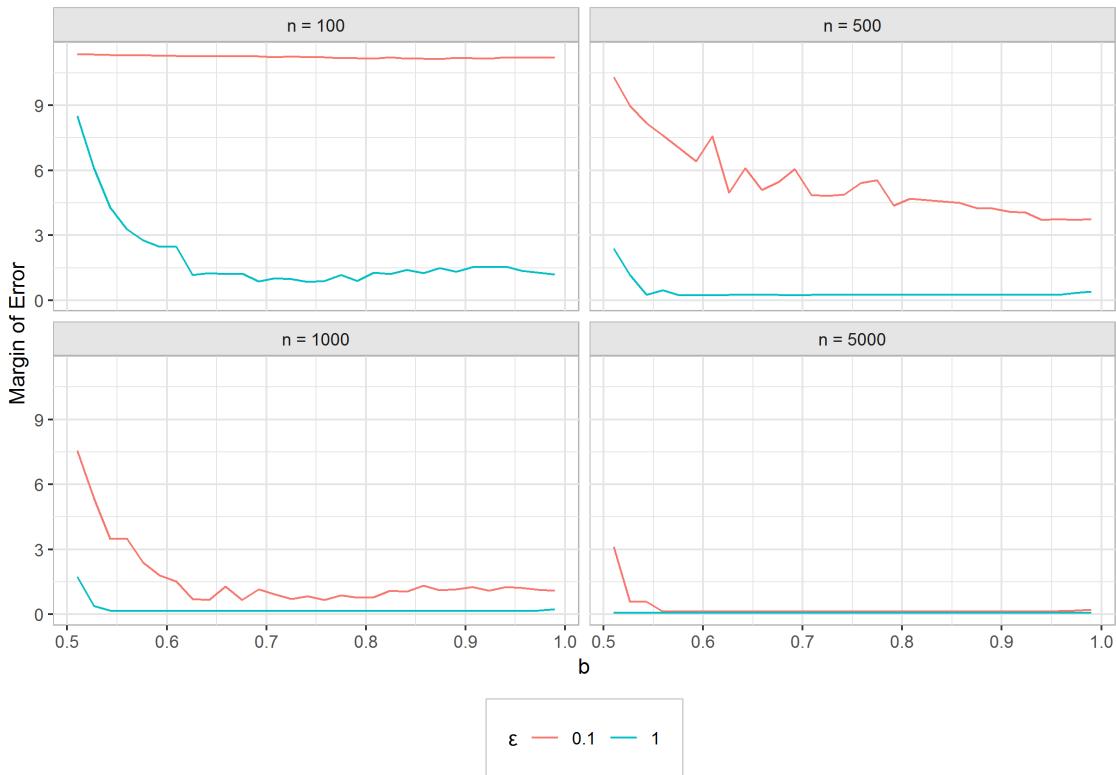


Figure 15: The performance of CENQ at various database sizes varying b , the percentile used for generating private measure of spread, with ε -allocation ρ set to be 0.5. CENQ tends to perform better when b falls into the range of approximately $(0.65, 0.8)$, where the MoE of confidence intervals tend to be relatively small for all database sizes. We here choose $b = 0.65$ as the optimized value of b for CENQ.

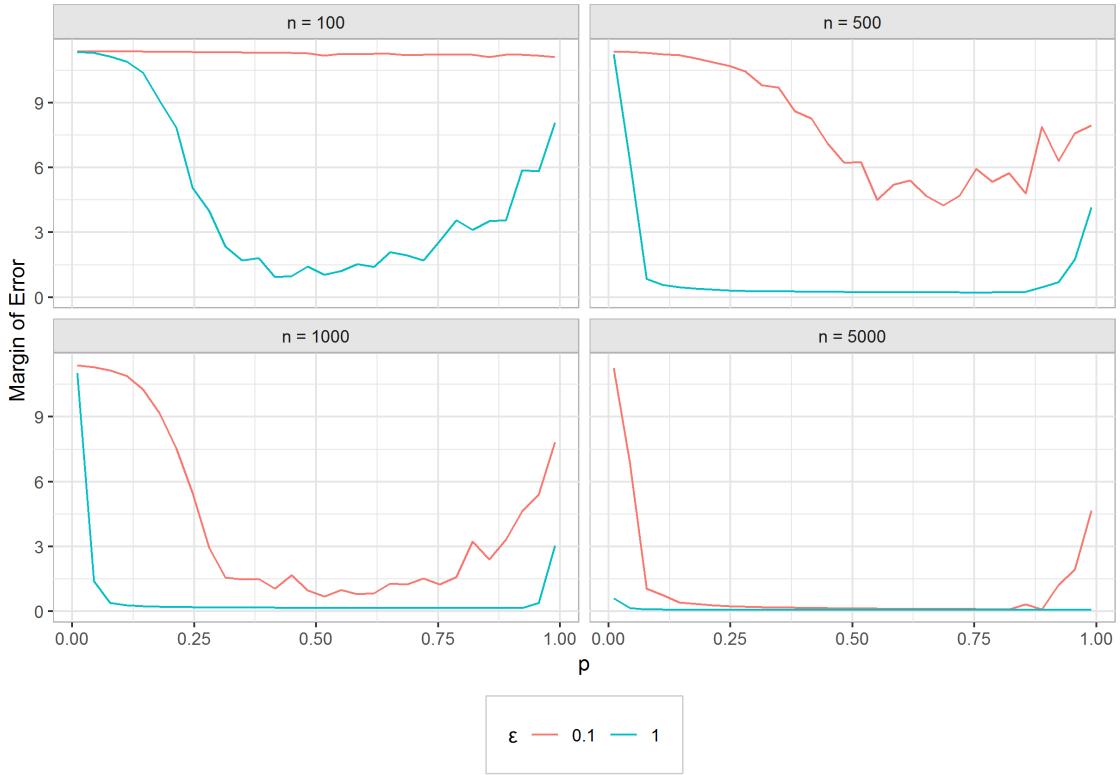


Figure 16: The performance of CENQ at various database sizes varying ε -allocation, with b set to be 0.65. CENQ tends to perform better with ρ in the range of approximately $(0.5, 0.75)$, where the MoE of the confidence intervals are small for all database sizes. We here choose $\rho = 0.5$ as the optimized ε -allocation.

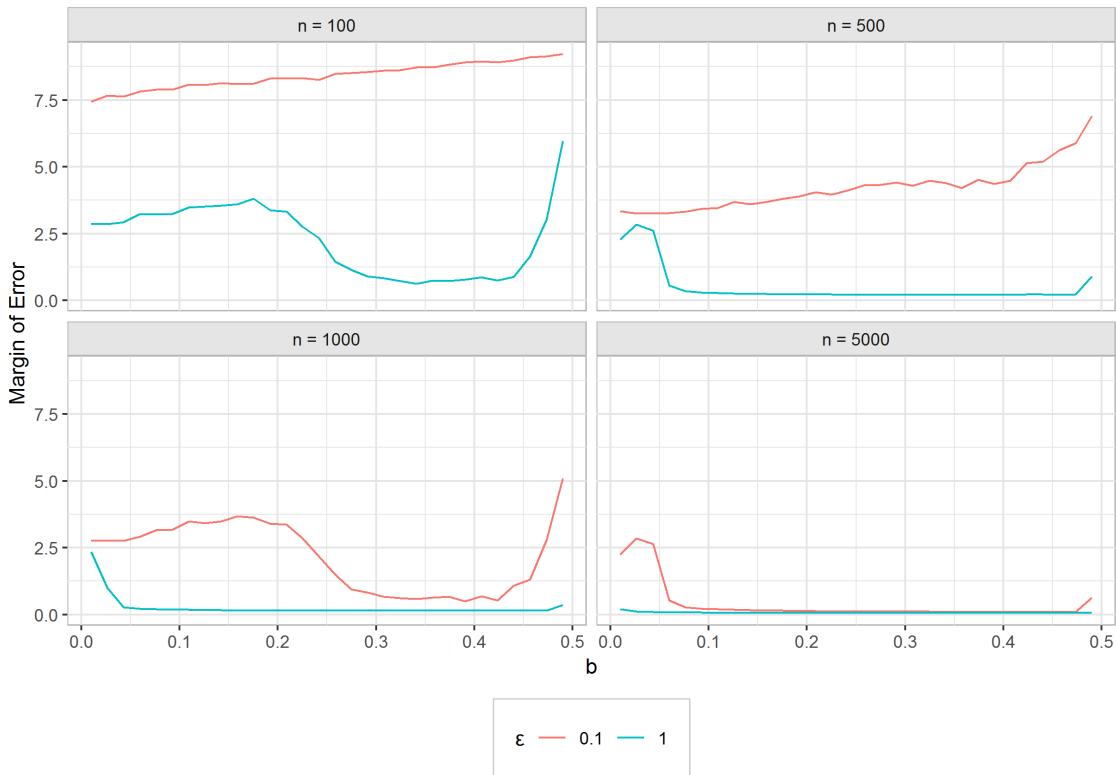


Figure 17: The performance of **SYMQ** at various database sizes varying b , the percentile used for generating private measure of spread. **SYMQ** tends to perform better when b falls into the range of approximately $(0.3, 0.45)$, where the MoE of confidence intervals are minimized for all database sizes. We here choose $b = 0.35$ as the optimized value of b for **SYMQ**.

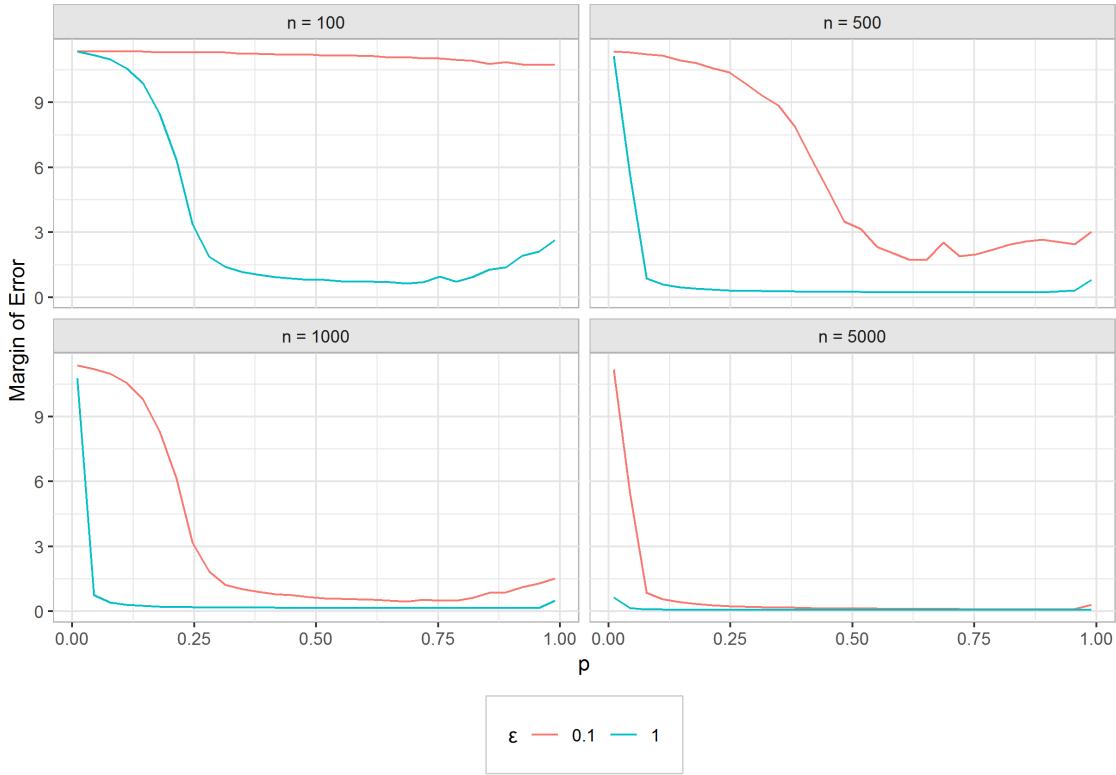


Figure 18: The performance of MOD at various database sizes varying ε -allocation. MOD tends to perform better when ρ falls into the range of approximately $(0.45, 0.55)$, where the MoE of confidence intervals are minimized for all database sizes. We here choose $\rho = 0.5$ as the optimized value of b for MOD.